8

ENCICLOPEDIA IBEROAMERICANA DE FILOSOFÍA

La mente humana

Edición de
Fernando Broncano

Editorial Trotta

Consejo Superior de Investigaciones Científicas

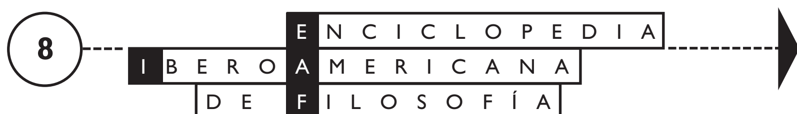


La mente humana

La mente humana

Edición de
Fernando Broncano

Editorial Trotta
Consejo Superior de Investigaciones Científicas



Catálogo general de publicaciones oficiales
http://publicacionesoficiales.boe.es/



EDITORIAL TROTTA

Primera edición: 1995
Primera reimpresión: 2007

Ferraz, 55. 28008 Madrid
Teléfono: 91 543 03 61
Fax: 91 543 14 88
E-mail: editorial@trotta.es
http://www.trotta.es

Departamento de Publicaciones
Vitruvio, 8. 28006 Madrid
Teléfono: 91 561 62 51
Fax: 91 561 48 51
E-mail: publ@orgc.csic.es

Diseño
Joaquín Gallego

ISBN: 978-84-87699-48-1 (Obra completa)
ISBN (edición digital pdf): 978-84-9879-393-2 (vol. 8)
NIPO: 653-07-035-5

Comité de Direc

Manuel Reyes Mate
Director del proyecto

León Olivé

Osvaldo Guariglia

Miguel A. Quintanilla

Pedro Pastur
Secretario administrativo

Comité Académico

Javier Muguerza	*Coordinador*
José Luis L. Aranguren	España
Ernesto Garzón Valdés	Argentina
Elías Díaz	España
Fernando Salmerón	México
Luis Villoro	México
Ezequiel de Olaso	Argentina
David Sobrevilla	Perú
Carlos Alchourrón	Argentina
Humberto Giannini	Chile
Guillermo Hoyos	Colombia
Javier Sasso	Venezuela

Instituciones académicas responsables

Instituto de Filosofía del C.S.I.C., Madrid.

Instituto de Investigaciones Filosóficas de la U.N.A.M., México (Directora Olbeth Hansberg).

Centro de Investigaciones Filosóficas, Buenos Aires (Director Mario Presas).

La Enciclopedia IberoAmericana de Filosofía es un proyecto de investigación y edición, puesto en marcha por el Instituto de Filosofía del Consejo Superior de Investigaciones Científicas (Madrid), el Instituto de Investigaciones Filosóficas de la Universidad Nacional Autónoma de México y del Centro de Investigaciones Filosóficas (Buenos Aires), y realizado por filósofos que tienen al español por instrumento lingüístico.

Existe una pujante y emprendedora comunidad filosófica hispanoparlante que carece, sin embargo, de una obra común que orqueste su plural riqueza y contribuya a su desarrollo. No se pretende aquí una enciclopedia de filosofía española sino articular la contribución de la comunidad hispanoparlante a la filosofía, sea mediante el desarrollo cualificado de temas filosóficos universales, sea desentrañando la modalidad de la recepción de esos temas filosóficos en nuestro ámbito lingüístico.

La voluntad del equipo responsable de integrar a todas las comunidades filosóficas de nuestra área lingüística, buscando no sólo la interdisciplinariedad sino también la internacionalidad en el tratamiento de los temas, nos ha llevado a un modelo específico de obra colectiva. No se trata de un diccionario de conceptos filosóficos ni de una enciclopedia ordenada alfabéticamente sino de una enciclopedia de temas monográficos selectos. La monografía temática permite un estudio diversificado, como diverso es el mundo de los filósofos que escriben en español.

La Enciclopedia IberoAmericana de Filosofía es el resultado editorial de un Proyecto de Investigación financiado por la Comisión Interministerial de Ciencia y Tecnología y por la Dirección General de Investigación Científica y Técnica del Ministerio de Educación y Ciencia. Cuenta también con la ayuda de la Consejería de Educación y Cultura de la Comunidad Autónoma de Madrid.

CONTENIDO

PRESENTACIÓN

Fernando Broncano

La filosofía de la mente es una de las disciplinas filosóficas que ha recibido mayor atención en los últimos años. Es también, o quizás por ello, una de las disciplinas donde se han producido novedades filosóficas que van más allá del mero comentario histórico y generan nuevas visiones del mundo. Como se comprobará leyendo este volumen, apenas hay historia de la filosofía explícita en él. No por desprecio o ignorancia, sino porque la marea de nuevos problemas a los que enfrentarse deja poco espacio para la filología. Es, sobre todo, un ámbito en el que las discusiones versan acerca de problemas vivos, importantes y últimos. Hay varias causas que lo explican. La primera de ellas es que la mente humana constituye una de las fronteras donde la investigación científica está haciendo los avances más espectaculares del siglo. El desarrollo es pluridisciplinar, rapidísimo, desordenado, desigual. Mucho más que en las ciencias maduras pero de lento desarrollo, las ciencias de la mente exigen imaginación, propuestas nuevas, discusiones vivas y un entusiasmo por el descubrimiento que atraen irresistiblemente al filósofo curioso por su mundo. La filosofía de la mente pertenece a la vieja tradición de la filosofía natural, la filosofía que especulaba no sobre, ni antes, ni después, ni debajo de las ciencias, sino al compás de ellas, revuelta con los científicos y despreocupada de las matrices disciplinares y las profesiones académicas. Quienes se dejan envolver por esta corriente tienen el privilegio de ser arrastrados por discusiones en las que suenan diferentes voces que hablan desde distintas disciplinas, con términos engañosamente iguales en ocasiones y sobre las mismas cuestiones de fondo bajo la aparente diversidad, mas siempre sobre problemas muy relevantes. La lingüística, la psicología, la inteligencia artificial, las ciencias cognitivas, la robótica, la biología evolucionista, la antropología, la etología, las neurociencias,

la neuropsicología, la neurofisiología, neurocomputación, la lógica y la teoría de la computabilidad, las matemáticas de los sistemas dinámicos, la filosofía del lenguaje. Toda aportación es bien recibida si tiene algo profundo que decir. La naturaleza de la mente ha dejado de ser el territorio exclusivo de los psicólogos; es un territorio abierto a la especulación.

No es casual, pues, que en la filosofía de la mente se hayan producido algunos de los más importantes desarrollos filosóficos del siglo. A pesar de que todavía no tienen un reconocimiento amplio en la comunidad filosófica, las diversas formas de funcionalismo son, por poner un ejemplo, la principal innovación metafísica desde las polémicas del XVIII y XIX, más allá de la encrucijada entre el materialismo y el idealismo. Como también son nuevas las preguntas que plantea el eliminativismo. Lo mismo podríamos decir de la noción de sobreveniencia, heredera del emergentismo de los años treinta, o de las nuevas perspectivas lógicas, no logicistas, sobre la naturaleza del razonamiento real de los sujetos reales.

Muchas teorías y problemas discutidos en este volumen parecerán esotéricos y lejanos a la intuición común. Es porque son teorías y problemas nuevos, técnicos, que exigen conceptos nuevos y técnicos, pero, sobre todo, porque modifican profundamente la intuición común que tenemos acerca de nuestra propia mente. Por ejemplo, la naturaleza del contenido mental, que afecta profundamente a la idea de idea, prácticamente intacta desde Platón, o las críticas a la vieja idea de la mente dividida en facultades, herencia aristotélica también intacta bajo tantos cambios filosóficos superficiales, o el cuestionamiento irreversible de la concepción neocartesiana, uno de los ejes centrales de la filosofía de nuestro siglo.

El volumen se ha concebido en el mismo espíritu que guió a los redactores de *L'Encyclopédie*. Se busca claridad expositiva, didáctica filosófica y amplitud de perspectivas, pero no conclusiones estables y definitivas. Nos hemos propuesto hacer un volumen histórico, que tenga sentido como un manual de intervención inmediata en la discusión actual, pero no un volumen para la historia. Hemos dividido el volumen en los temas esenciales desde la perspectiva contemporánea, esto es, los temas que son objeto de mayor discusión entre los filósofos y los científicos.

La primera parte, que abarca los artículos de Eduardo Rabossi, Manuel García Carpintero, Daniel Quesada, Jesús Ezquerro y Josep Corbí y Josep Lluís Prades, se ocupa de la introducción a las concepciones sobre la mente vigentes en la actualidad. Se examina el problema de cómo entender el fenómeno de lo mental desde una perspectiva científica y naturalista. Eduardo Rabossi estudia las tesis de la identidad entre fenómenos mentales y fenómenos cerebrales, una alternativa que adopta una posición muy clara y operativa para el trabajo científico, pero que ha sufrido

en las dos últimas décadas el ataque de los argumentos del funcionalismo, quizás la más importante de las concepciones de lo mental en este siglo, puesto que permite recoger todo lo que el dualismo mente-cuerpo tradicional podía explicar sin caer en su principal dificultad: cómo entender el hecho, difícilmente discutible, de la interacción entre fenómenos físicos y mentales, sin suponer algo así como la armonía preestablecida entre lo que piensa la mente y lo que hace el cuerpo. Este problema es el que ha hecho que alternativas neocartesianas, como la representada por Popper, no hayan tenido buena acogida. Manuel García Carpintero reconstruye el concepto y la historia del funcionalismo comparándola con sus más directos rivales, el conductismo y la concepción cartesiana de lo mental. Daniel Quesada, por su parte, escribe sobre una forma especial de funcionalismo, el llamado funcionalismo biológico, una de las más novedosas concepciones de lo mental, que recupera la naturalidad de lo mental en un mundo de funciones biológicas que tienen una explicación diferente a la de los fenómenos fisicoquímicos de los órganos que las realizan. El funcionalismo biológico nos recuerda que la mayoría de los problemas filosóficos que encontramos en la relación entre fenómenos mentales y fenómenos corporales los encontramos también cuando examinamos las relaciones entre órganos y las funciones que cumplen las actividades que realizan esos órganos. Jesús Ezquerro, en su trabajo sobre las arquitecturas de lo mental, nos ofrece una idea fidedigna de las aportaciones que está haciendo la llamada inteligencia artificial a la comprensión de la naturaleza de los estados mentales. Después de varias décadas de discusiones acerca de la metáfora del ordenador, ha ido imponiéndose la realidad de la investigación sobre las discusiones de principio. No importa tanto si los ordenadores y los cerebros son sistemas que pertenecen a una clase próxima o igual, cuanto que la metáfora o el modelo ha sido muy productiva como fuente de imaginación creadora y ha dado lugar a hipótesis audaces sobre el diseño de las funciones intelectuales superiores. Estas hipótesis son teóricas, en el sentido de que transcienden los datos psicológicos, ni siquiera se comprometen con el prejuicio de que vayan a ser inequívocamente humanas, pero las mejores tentativas que tenemos de aproximarnos a la estructura de las operaciones mentales de una manera científica. Puede hablarse de un antes y un después de las teorías de la arquitectura de lo mental. Ya comienza a ser normal en los contextos de psicología el uso de los esotéricos nombres de las teorías presentadas en este trabajo, manejadas como auténticas hipótesis psicológicas, sin importar demasiado si su origen fue el diseño de ordenadores inteligentes o la observación de la conducta de los propios hijos, como hacía Piaget. El trabajo de Jesús Ezquerro no es, sin embargo, solamente informativo, sino que se enfrenta a los problemas filosóficos, genuinamente filosóficos, que presenta esta idea de estudiar la mente como un edificio del que puede extraerse el plano en sucesivos ni-

veles de aproximación. Josep Corbí y Josep Lluís Prades abordan la cuestión de lo que podríamos llamar la «arquitectura» de lo mental desde la perspectiva que se ha denominado «conexionista». El conexionismo es la venganza contra lo que durante las últimas décadas se denominó la «metáfora del ordenador» como imagen de la naturaleza de lo mental. Tras el cerebro como un ordenador, la inteligencia artificial ha pasado a servirse de las neuronas de verdad como modelo para construir ordenadores y programas que se parezcan al cerebro en su estructura física y funciones. Como resultado, se han elaborado espectaculares programas que son capaces de aprender a realizar tareas antes exclusivas de los seres humanos. El conexionismo es la filosofía que se propone estudiar la mente usando estos sistemas como modelos de las principales funciones mentales y es uno de los pocos casos en los que realmente podemos hablar de cambio de paradigma en el estudio de un fenómeno humano. En este trabajo se presentan las principales regiones a las que afectaría la revolución conexionista, de extenderse y tener éxito.

Una segunda parte se ocupa del estudio de aquello que desde Brentano se ha concebido como la característica definitoria de lo mental, la propiedad que denotamos con el nombre de *intencionalidad*, el estar «dirigido a» un cierto objeto. Juan José Acero ha sido el encargado de presentar las principales concepciones del aspecto más importante de los fenómenos intencionales, el que tengan contenido mental. Cuál sea la naturaleza de este contenido y cómo acomodar su estudio en la empresa científica es el objetivo de su trabajo. En él se presenta un panorama completo de las teorías vigentes sobre el contenido mental que tienen vigencia operativa en las ciencias que se ocupan de los fenómenos mentales. Manuel Liz tiene sobre sí la tarea de presentar uno de los problemas más complicados que uno puede encontrar cuando se enfrenta al estudio de la naturaleza de lo mental. Se trata de la cuestión de cómo los contenidos mentales, a los que acudimos para dar razón de nuestra conducta intencionada e intencional, pueden causar hechos cuya naturaleza no es mental. Si los contenidos son cosas físicas, como piensan quienes identifican estados mentales con procesos cerebrales, entonces no habrá problema, porque ocurrirá que unos procesos físicos causan otros procesos físicos. Pero esto es lo mismo que decir que los contenidos mentales no son nada, que no los necesitamos para explicar la conducta, porque nos basta, o nos bastará cuando se complete, la descripción neuroquímica del cerebro. De manera que llegaríamos a concluir que si los conceptos mentales cumplen su función de ayudar a explicar nuestra conducta, entonces no son necesarios, y si no cumplen su función entonces sí que son realmente innecesarios. Liz nos introduce al concepto filosófico de sobreveniencia o superveniencia, quizás uno de los pocos conceptos realmente nuevos que han surgido en la filosofía de este siglo. Josefa Toribio trata un tema relacionado con la discusión anterior y derivado del

desarrollo de las ciencias que se ocupan de lo mental: ¿hasta qué punto está bien fundamentado todo este aparato conceptual con el que describimos los fenómenos mentales? Sabemos por la historia de la ciencia y de la cultura que nuestros conceptos sobre el universo han cambiado de forma radical a lo largo del tiempo, y aunque no haya cambiado nuestro modo de hablar sobre las cosas, estamos convencidos de que la mayoría de las veces no es más que un modo de hablar. Decimos que sale el sol o que el calor pasa de los cuerpos calientes a los cuerpos fríos, pero sabemos bien que estas expresiones sirven sólo para andar por casa y no para hacer física. Sin embargo, los conceptos con los que describimos lo mental, la idea de idea pongamos por caso, no solamente no se ha modificado desde los más remotos tiempos, sino que muchos piensan que no se puede ni debe modificar. En el lado contrario, algunos filósofos tienen una opinión muy diferente y han calificado a toda esta terminología de «psicología popular» (*folk*, en el término inglés original), con el objetivo de asimilarla a las charlas que uno escucha en el mercado cuando oye describir a la gente sus enfermedades y que bien pueden ser calificadas de medicina popular. A esta concepción se la denomina eliminativismo y, como su propio nombre indica, propone una solución realmente final para el problema mente-cuerpo.

La tercera parte aborda algunos de los principales aspectos de lo que tradicionalmente han sido consideradas facultades de la mente: lenguaje, pensamiento, racionalidad, percepción, conciencia. Se han abordado, lo mismo que los temas anteriores, desde la perspectiva de su último tratamiento. Antoni Gomila se ocupa del lenguaje y de cómo pudo emerger en el tiempo esta facultad, la más importante para la configuración de la mente humana. Ya nadie duda a estas alturas de que el *homo sapiens sapiens* es el único mono gramático que subsiste en la superficie de la Tierra y de que la capacidad para el lenguaje es en gran parte innata, y de que la cuestión de cómo tal cosa pudo ser posible es quizás una de las preguntas más profundas que uno puede dirigir a la teoría de la evolución. El estudio evolutivo del lenguaje no es solamente un problema histórico o antropológico: aporta también una forma de mirar el propio lenguaje, pero sobre todo una forma de mirar cuál es nuestra identidad natural como especie. Fernando Broncano desarrolla varios de los aspectos correlacionados con lo que tradicionalmente hemos llamado razón o racionalidad: qué modelo describe la racionalidad, qué relaciones hay entre racionalidad y emociones, si puede hablarse individualmente de racionalidad o si acaso la racionalidad no sería una herencia de nuestra constitución como especie social, la herencia de los monos que se organizaron en bandas y que, al hacerlo, generaron un nuevo medio ambiente lleno de nuevos problemas que nacen de las relaciones sociales y no de la exigencia de supervivencia física. Vicente Sanfélix presenta las principales concepciones actuales sobre la percepción, otra de las grandes regio-

nes de investigación de lo mental. Alfonso García Suárez nos muestra precisamente un problema general de todo lo mental, pero en el que la percepción se involucra de modo especial, a saber, el problema de las cualidades fenomenológicas o «qualia», de hecho el lugar donde nació el problema de lo mental, cuando Descartes y sus contemporáneos notaron para el resto de los tiempos que no hay cosas rojas ni amargas, sino percepciones de algo como rojo o como amargo. García Suárez nos introduce a la panoplia de argumentos y concepciones sobre la realidad de estas propiedades que tradicionalmente se llamaban secundarias, y con ello a una de las cuestiones de más difícil tratamiento de la filosofía de la mente, pero también de las piedras angulares de cualquier teoría de lo mental. Por último, Enrique Villanueva nos presenta el problema de la conciencia, la facultad que casi todos relacionaríamos de manera más inmediata con la naturaleza de lo mental. ¿Es soluble el problema de explicar la naturaleza de la conciencia? ¿es siquiera planteable? ¿cuántas y cuáles son las dimensiones del problema? Muchos son los que creen que la conciencia es el muro contra el que chocarán todos los intentos de integrar nuestra naturaleza en la naturaleza de las cosas. Otros, que bajo el nombre de conciencia agrupamos demasiadas cosas, no todas ellas homogéneas.

Ninguna de las discusiones, de los argumentos y teorías resulta fácil de seguir. Aparecen muchos nombres nuevos que seguramente no son familiares y teorías que en apariencia resultan ajenas a la filosofía tradicional. Pero sólo en apariencia. Hay una continuidad en el desarrollo del estudio de la mente desde los primeros tiempos de la filosofía. Pero del mismo modo que después de la revolución científica aparecen nuevos conceptos que determinan el modo de pensar la naturaleza, en el siglo XX la mente es la frontera en la que se producen los cambios más profundos en la concepción de nuestra propia identidad. Las ciencias de la mente son ciencias naturales, pero quién se atrevería a negarles el estatuto de ciencias humanas y, a la inversa, son ciencias humanas a las que solamente unos cuantos se niegan a reconocer como naturales.

¿Qué somos? ¿de dónde venimos? ¿qué nos cabe esperar? ¿qué es el hombre?: que nadie espere una respuesta a estas preguntas leyendo este volumen. Pero que nadie crea que se puede responder a ellas sin responder a otras cuestiones que aquí se suscitan.

LA TESIS DE LA IDENTIDAD MENTE-CUERPO

Eduardo Rabossi

I. EL PROBLEMA MENTE-CUERPO Y LA TESIS DE LA IDENTIDAD

El problema mente-cuerpo incluye, básicamente, tres temas enigmáticos, íntimamente relacionados entre sí: 1. la *naturaleza* de los fenómenos mentales *vis-a-vis* la naturaleza de los fenómenos físicos (corporales) (¿qué es lo que caracteriza de manera esencial a los fenómenos mentales?); 2. el *status ontológico* de los fenómenos mentales (¿constituye lo mental un ámbito propio de la realidad, distinto del ámbito físico?), y 3. la índole de la eventual *relación* de los fenómenos mentales con los fenómenos físicos (¿es una relación de tipo causal?; si no lo es, ¿cuáles son sus características típicas?). A lo largo de los siglos, muchos filósofos han intentado dar respuesta a estos enigmas. La insistencia está justificada. Las cuestiones que involucra el problema mente-cuerpo tienen en sí mismas una gran importancia teórica; además, la solución global del problema posee una significación especial. Un enfoque filosófico adecuado de la persona humana exige, entre otras cosas, estar en condiciones de dar una respuesta al problema de la mente y el cuerpo.

En este tema, como en tantos otros, la viabilidad de una teoría filosófica no se puede medir únicamente en términos de su coherencia interna. En el caso del problema mente-cuerpo hay dos elementos extrafilosóficos que juegan un papel importante a la hora de juzgar la adecuación de una teoría: las convicciones de sentido común acerca de la mente y del cuerpo, y la información que proporcionan ciertas disciplinas científicas.

Nuestras relaciones interpersonales y la concepción que tenemos de nosotros mismos como personas están mediadas por un conjunto de convicciones básicas que involucran lo mental y lo corporal. Así,

— nos reconocemos como sujetos legítimos de propiedades físicas (corporales) (por ejemplo, medir 1,70 m., pesar 70 kg., ser calvo) y de propiedades mentales o psicológicas (por ejemplo, desear comer un buen cocido, creer que el Teide es el monte más elevado de España, sentir dolor de muelas),

— consideramos que las propiedades mentales son de una índole diferente a la de las propiedades fisicas, pues no parecen ser simples manifestaciones de la materia física, o estar constituidas meramente por materia física,

— consideramos, también, que cada persona tiene una mente que le es «propia» y que la constituye como la persona que es,

— pero no concebimos a las propiedades mentales ni a las mentes como totalmente independientes del ámbito físico, porque sabemos que la actividad cerebral es una condición necesaria de nuestra vida psíquica, que nuestra mente está donde está nuestro cuerpo, y que las aptitudes cognitivas que nos diferencian de otros seres vivos se deben, en gran medida, a nuestro desarrollo anatómico y neurológico,

— reconocemos, asimismo, que las propiedades mentales tienen eficacia respecto de las propiedades físicas, y viceversa (atribuyo el dolor que experimento a una infección en mi molar inferior izquierdo, relaciono mi creencia con el hecho de que el Teide tenga cierta altura, y mi deseo de comer un buen cocido explica por qué camino en dirección a la Fonda del Arco).

Es interesante advertir que estas convicciones de sentido común exhiben una tensión interna que tiene una contrapartida en la reflexión filosófica acerca del problema mente-cuerpo: de un lado, los fenómenos mentales, la mente, no parecen ser físicos, pues, en un sentido, son independientes de los fenómenos corporales; del otro lado, tienen que ser físicos porque, en un sentido distinto, dependen de lo corporal [1].

Como he indicado más arriba, hay otro elemento que el filósofo de la mente debe tomar en cuenta al encarar el problema mente-cuerpo: la información que proviene de las disciplinas científicas que investigan aspectos de la psique, el cuerpo, y sus eventuales conexiones. La biología, la neurociencia, la psicología, la neuropsicología y, más recientemente, la ciencia cognitiva, son las disciplinas pertinentes. Así como una teoría filosófica adecuada de la mente debe dar cuenta de las convicciones de sentido común (pues aun su rechazo exige ofrecer razones adecuadas), también tiene que ser sensible a los hallazgos científicos (pues aun cuando se practique la filosofía con una metodología puramente *a priori*, una con-

1. McGinn (1982).

dición mínima a satisfacer es que sus conclusiones no resulten contradictorias con tales hallazgos)[2].

La tesis de la identidad mente-cuerpo (también denominada teoría de la identidad mente-cuerpo, teoría de la identidad mente-cerebro, teoría de la identidad psico-física, teoría materialista de la identidad, teoría de la identidad de la mente, materialismo de estado [del sistema nervioso] central, materialismo reductivo, fisicalismo de tipos y, a veces, simplemente, materialismo, fisicalismo o teoría de la identidad) constituye una respuesta radical a los enigmas del problema mente-cuerpo. En términos generales, la tesis de la Identidad (TI, en lo sucesivo) sostiene que los fenómenos mentales *son* (numéricamente idénticos a) estados físicos (estados cerebrales) y, *a fortiori*, que la mente *es* (numéricamente idéntica a) el cerebro. Planteada en estos términos, la TI no se diferencia de un materialismo tosco, a la manera de ciertas versiones producidas en el siglo XIX. Pero la TI es mucho más sofisticada de lo que esa primera caracterización pueda llegar a sugerir. Su surgimiento está asociado a un factor extrafilosófico: el desarrollo de la neurofisiología debido a E. D. Adrian, W. Penfield, D. O. Hebb y W. S. McCullock, entre otros, y al avance de la biología molecular[3]; y a un factor filosófico: la consiguiente convicción de que contamos con elementos suficientes como para pensar que los organismos pueden ser considerados mecanismos físico-químicos y que la conducta de los seres humanos pueda ser explicable algún día en términos de ese tipo de mecanismos[4]. Los defensores de la TI se proponen ofrecer un marco teórico que resuelva los enigmas del problema mente-cuerpo, en franca oposición a las ofertas teóricas del dualismo cartesiano y del conductismo (aunque incorporando algunas tesis de estas posiciones).

Los trabajos seminales de la TI se deben a los filósofos australianos U. T. Place y J. J. C. Smart, y al filósofo norteamericano H. Feigl. La obra de Smart fue la que ejerció la influencia mayor[5]. La TI ocupó el centro de la discusión filosófica durante la década de los años sesenta. El funcionalismo, cuyas primeras formulaciones se producen a fines de esa década, borró a la TI del escenario filosófico. Desde mediados de la década de los setenta, la gran mayoría de los filósofos de la mente considera que la TI es un intento importante pero fallido de dar respuesta a los enigmas del problema mente-cuerpo.

En las páginas que siguen, presentaré la matriz teórica inicial de la TI

2. Sobre el problema mente-cuerpo y las teorías filosóficas que intentan darle solución, véase Campbell (1970), Bunge (1980) y Churchland (1988).

3. Feigl (1960), Armstrong (1965).

4. Smart (1959).

5. Place (1956) (1960), Smart (1959) (1961) (1962) (1963), Feigl (1958) (1960). Sobre el desarrollo de la TI en Australia, véase la Introducción a Presley (1967). Las siguientes compilaciones recogen los trabajos y las discusiones más importantes: Presley (1967), Borst (1970), Rosenthal (1971).

(Sección II) y algunas de las críticas más importantes que se le pueden formular (Sección III). Desarrollaré luego dos matrices teóricas que intentan superar las limitaciones de la versión original: el eliminativismo y la teoría de la identidad de rol causal (Secciones IV y V). Concluiré con una referencia al argumento de la realizabilidad variable y al reduccionismo (Sección VI). No me propongo historiar los avatares de la TI. Me interesa mostrar los rasgos básicos de la dialéctica teórico-conceptual a que da lugar la defensa y la crítica de una posición radical como la TI.

Debo confesar mi simpatía por la TI y por algunas de las convicciones filosóficas que la motivan, así como mis reservas acerca de la viabilidad del funcionalismo. Dada la índole del presente trabajo, no entraré en esas cuestiones polémicas. La Sección VI recoge mínimamente algunas inquietudes al respecto.

II. LA MATRIZ TEÓRICA INICIAL DE LA TESIS DE LA IDENTIDAD

Existen excelentes razones filosóficas para rechazar el dualismo y el conductismo como soluciones globales al problema mente-cuerpo.

El *dualista* sostiene que los fenómenos mentales son radicalmente distintos de los fenómenos físicos porque sus propiedades esenciales son diferentes. La no-espacialidad, la privacidad, la introspectibilidad, la intencionalidad, el carácter interno y/o la conciencia, caracterizan a los fenómenos mentales. La espacialidad y el carácter público son las propiedades privativas de los fenómenos físicos. Pero ¿cuál es el fundamento de esa diferencia crucial? El *dualista substancialista* funda la diferencia en una circunstancia ontológica: cada tipo de fenómeno y las propiedades esenciales que lo caracterizan, son la manifestación de una substancia o de un tipo de substancia peculiar. Esto implica, entre otras cosas, postular un hiato definitivo entre los dos ámbitos y, consiguientemente, rechazar toda estrategia que se proponga definir o reducir uno al otro. Pero, dado el hecho obvio de que lo mental y lo físico se influyen mutuamente, el dualista se ve obligado a postular algún tipo de relación entre ellos. Interacción, paralelismo, armonía pre-establecida, epifenomenalidad, son algunas de las hipótesis explicativas intentadas por los dualistas.

El dualismo es insostenible por varias razones. Las más evidentes son que 1. no tiene manera de explicar cabalmente cómo se relacionan los ámbitos que postula; 2. la postulación a ambos ámbitos sólo obedece a una estrategia ontológica *ad hoc*; 3. carece de criterios autónomos de identidad para las substancias mentales, pues tiene que apelar a criterios de carácter público o caer en el solipsismo; 4. es incompatible con ciertos principios básicos de la ciencia física, en particular con el principio de conservación de la energía; 5. torna impensable una ciencia de lo mental

integrada al cuerpo de la ciencia; 6. se limita a recoger ciertos rasgos atribuidos precríticamente a los fenómenos mentales, sin indagar en sus mecanismos ni explicar por qué tienen la pertinencia teórica que les atribuye.

El dualismo debe ser rechazado, pero es conveniente advertir que, al rechazarlo en base a *ese tipo de argumentos*, se debe estar dispuesto a defender un monismo ontológico (sólo son válidas las entidades físicas), a admitir que los fenómenos mentales son definibles en términos de fenómenos físicos o reducibles a ellos, a pensar que las ciencias constituyen un continuo conceptual y legal que supone ciencias de base (en particular, la física) y a sostener que los rasgos que atribuimos precríticamente a los fenómenos mentales pueden ser acomodados, de alguna manera, dentro de un marco no dualista.

El *conductismo* plantea una fuerte crítica al dualismo, que no se inspira en un punto de vista científico, sino en los cánones del análisis conceptual. Su tesis básica es que, en principio, los términos mentales y, por ende, las oraciones en los que figuran de manera esencial, pueden ser transcriptos sin pérdida de significado en términos y oraciones acerca de la conducta y de ciertas circunstancias antecedentes observables. No se requiere que la conducta sea expresa. Las transcripciones apelan a disposiciones de los agentes a producir un cierto patrón de conducta (una cierta respuesta conductual, actual o posible) ante ciertos tipos de estímulos del medio ambiente. Todo esto implica negar la pertinencia causal de los estados y procesos internos de los agentes y la consiguiente necesidad de incluirlos en un modelo filosófico de la vida mental. Los estados mentales son estados conductuales o disposiciones a producirlos. El vocabulario mental enraíza su significado en la conducta pertinente, por lo que la referencia a causas internas, fisiológicas o neurológicas, carece de importancia filosófica.

El conductismo también es insostenible. Algunas razones obvias son que 1. las prometidas transcripciones resultan demasiado generales y, en consecuencia, dudosamente elucidatorias, o bien implican un número no acotable de circunstancias específicas; 2. los análisis de tipo conductista sólo son posibles si se admiten ciertos supuestos acerca de los estados mentales de los agentes; 3. dos agentes pueden diferir en sus estados psicológicos pese a la similitud de sus respuestas conductuales; 4. hay estados mentales internos no acompañados por tales respuestas; 5. no resulta admisible negar la relevancia de la ciencia respecto del problema mente-cuerpo.

Si se rechaza el conductismo en base a *ese tipo de argumentos*, se debe estar dispuesto a sostener la relevancia de los estados internos de los agentes y sus conexiones causales, a dudar del interés teórico de los planteos filosóficos que se centran en el análisis del significado de los términos y oraciones mentales, a explicitar en qué sentido los hallazgos científicos

resultan relevantes para un enfoque filosófico de la mente y el cuerpo, y a ubicar la realidad de los fenómenos mentales dentro de ese marco.

Este breve paso por los vericuetos teóricos del dualismo y el conductismo resulta indispensable para delinear el trasfondo filosófico de la TI. *Si* el dualismo y el conductismo resultan ser inviables como respuestas globales al problema mente-cuerpo, *si* se asumen las consecuencias de los argumentos que justifican su rechazo, *si* se reconoce la relevancia de ciertos desarrollos científicos, *si* se tiene, además, la convicción de que la conducta de los seres humanos va a ser explicable algún día en base a mecanismos físico-químicos, es posible concebir una respuesta al problema mente-cuerpo que sea novedosa y radical *vis* a *vis* los intentos conocidos. La TI es, según sus defensores, tal respuesta.

La matriz teórica de la versión inicial de la TI (TII en lo sucesivo) puede esquematizarse en los siguientes términos:

(A) Los fenómenos mentales son fenómenos *internos* de los seres humanos. En este punto la TII concuerda con el dualismo y discrepa con el conductismo.

(B) Los fenómenos mentales son idénticos a estados neurológicos del sistema nervioso central. *El dolor* es numéricamente idéntico a *Disparos de las fibras-c*, digamos. La identidad debe entenderse en un sentido estricto. No se sostiene que *El dolor* sea espacial o temporalmente contiguo a *Disparos de las fibras-c*, sino que *El dolor* es *Disparos de las fibras-c*. En este punto la TII discrepa con el dualismo y el conductismo. (Las cursivas con mayúscula indican que hablamos de los fenómenos, no de las palabras, y de los fenómenos tipo, no de sus casos; ver III.3, más adelante).

(C) Los enunciados que aseveran la identidad de los fenómenos mentales con estados neurológicos, expresan verdades contingentes. Son del mismo tipo que las verdades sintéticas y *a posteriori* expresadas, por ejemplo, por «Las nubes son conjuntos de gotas diminutas de agua», «El agua es H_2O», «El Lucero Matutino es el Lucero Vespertino», «Los genes son segmentos de moléculas de ADN», «El calor es energía cinética media». El desarrollo de la neurofisiología va a validar la identidad de los distintos fenómenos mentales tipo con los estados neurales tipo que les correspondan.

(D) El carácter contingente de tales identidades tiene consecuencias adicionales que el defensor de la TII está dispuesto a asumir. Es posible que las teorías neurofisiológicas cambien, que su desarrollo demuestre que la TII es científicamente inviable, que haya fenómenos mentales que no corresponden a estados neurofisiológicos y estados neurofisiológicos que, en definitiva, no sean correlacionables con determinados fenómenos mentales. La TII concuerda con el dualismo substancialista cartesiano en reconocer la posibilidad de estos dos últimos casos.

(E) La TII no es una tesis acerca del significado de los términos mentales. No sostiene que «El dolor» signifique lo mismo que «Disparos de las fibras-c», sino que cuando se reporta un dolor, por ejemplo, se reporta un proceso que *es* un estado cerebral. En este punto la TII discrepa nuevamente con el conductismo.

(F) Los fenómenos mentales están causalmente ligados entre sí y con situaciones estímulo del medio ambiente. Dada la identidad de los fenómenos mentales con estados cerebrales, la noción de causalidad involucrada es la noción *standard*, y los nexos causales postulados corresponden a nexos causales normales. La causalidad mental no es misteriosa, como en el dualismo.

(G) La equiparación de «El dolor es disparos de fibras-c» con «Los genes son segmentos de moléculas de ADN», por ejemplo, lleva a pensar en la posibilidad de reducir la psicología a la neurofisiología. En un proceso reductivo tal, la neurofisiología será la teoría de base para la reducción de la psicología. Las leyes puente establecerán las identidades entre las propiedades mentales y las propiedades neurológicas. La reducción permitirá fundar, en definitiva, las identidades empíricas postuladas. La posibilidad de la reducción se basa en una hipótesis empírica de carácter general: la unidad de la ciencia.

(H) La TII no sólo supone la esperanza razonable de que algún día la ciencia llegará a explicar la conducta de los seres humanos en base a mecanismos físico-químicos, sino que alega en su favor razones de parsimonia: el compromiso ontológico que asume la TII es menor que el que asume el dualismo. Por otra parte, la formulación de la TII no sólo prueba que no estamos obligados a dar una respuesta dualista al problema mente-cuerpo, sino que es posible darle una respuesta no dualista lógicamente coherente.

Cabe aclarar que originariamente Place y Smart restringieron el ámbito de la TII a las sensaciones. Su idea fue que el conductismo ofrece un tratamiento adecuado de nociones tales como conocer, creer, comprender, recordar, querer y tener intención, pero que el análisis conductista de las nociones de conciencia, experiencia, sensación e imagen es inadecuado. La razón es que los fenómenos denotados por este último grupo de nociones suponen el acaecimiento de un cierto proceso interno. Esta restricción en el alcance de la TII fue luego dejada a un lado.

III. ALGUNAS CRÍTICAS A LA MATRIZ TEÓRICA INICIAL DE LA TESIS DE LA IDENTIDAD

La TII produjo una reacción crítica en cadena. Durante una década y media ocupó el centro del escenario filosófico e incitó a una discusión viva y fructífera que sacudió los supuestos tradicionales de la filo-

sofía de la mente. No voy a exponer los detalles completos de la discusión[6]. Presentaré cuatro temas que plantean a la TII dificultades importantes.

1. *La neutralidad tópica*

Es posible dar una versión fregeana de las oraciones de identidad que postula la TII[7]. En «El dolor es Disparos de las fibras-c», por ejemplo, las expresiones «El dolor» y «Disparos de las fibras-c» tienen el mismo *denotatum* (la misma referencia), a saber, un cierto estado neurofisiológico. Pero su sentido (su significado) es distinto; es decir, el modo en que cada expresión presenta el *denotatum* (el criterio que supone) es diferente, así como son diferentes los modos de confirmación de los enunciados que los contienen. La versión fregeana evita, en principio, la tentación de argumentar que, porque dos tipos de expresiones (las fenoménicas y las neurofisiológicas) poseen «lógicas» diferentes, deberían denotar tipos de entidades ontológicamente diferentes. Pero, al mismo tiempo, pone en evidencia un problema general que afecta a las identidades que postula la TII.

Así como «El Lucero Matutino es el Lucero Vespertino» (el ejemplo dilecto de Frege) involucra que las propiedades que caracterizan al primero son lógicamente diferentes de las que caracterizan al segundo, en las identidades que postula la TII las propiedades que caracterizan a lo fenoménico son lógicamente diferentes de las propiedades que caracterizan a lo neurofisiológico. Más aún, ese *tiene* que ser el caso, porque una condición para que se puedan formular tales identidades, es que exista alguna propiedad fenoménica no poseída por el proceso neurofisiológico. Si no fuera así, las identidades no podrían siquiera pergeñarse. La conclusión es obvia: en las identidades postuladas por la TII hay un residuo tópico no eliminable.

Esta objeción se podría responder si se lograra reformular la parte mentalista (izquierda) de la identidad, conservando la referencia *vía* un modo de presentación neutral. La estrategia preferida[8] consiste, primero, en dejar de hablar de «objetos» fenoménicos (*El dolor, Las sensaciones*) con los que los agentes estarían relacionados, y preferir los reportes de experiencias fenoménicas («Veo una postimagen anaranjada», por ejemplo), negando que en un sentido estricto esos reportes involucren un compromiso con propiedades específicas. El segundo paso consiste en proponer una forma normal, tópicamente neutral (es decir, neutral respecto de compromisos materialistas o dualistas), que recoja lo que se dice

6. Véase al respecto Smart (1959), Borst (1970), Carruthers (1991) y MacDonald (1989).
7. Feigl (1958).
8. Smart (1959) (1960).

en el lenguaje corriente cuando se reporta una experiencia fenoménica. La tesis es que cuando alguien dice, por ejemplo,

(1) «Veo una postimagen anaranjada»,

realmente dice

(2) «*Algo acaece que es como lo que acaece* cuando tengo mis ojos abiertos y hay una naranja bien iluminada frente a mí»,

(en donde «Tengo los ojos abiertos y hay una naranja bien iluminada frente a mí» describe condiciones del agente y un estímulo físico). Se reconoce que la transcripción tópicamente neutral no especifica en qué respecto lo que acaece es similar (o no) a lo que acaece cuando... Esa falta de especificidad se considera una ventaja. La transcripción propuesta sólo supone que poseemos la aptitud de dar cuenta de que algo es como otra cosa, sin que podamos especificar en qué respecto lo es.

La crítica más contundente a esta defensa de la TII es la siguiente: (1) y (2) tienen el mismo significado (esto resulta de la afirmación de que cuando alguien dice (1) realmente dice (2). De la circunstancia de que (1) y (2) tengan el mismo significado se sigue que pueden sustituirse mutuamente. Concedamos que (1) es *siempre* sustituible por (2), esto es que cuando alguien dice (1), dice (2). Pero la conversa no se da: (2) no siempre es sustituible por (1). Supongamos que alguien dice

(3) Veo una forma redonda.

(3) no es sinónima de (1) y, sin embargo, *según sea la experiencia del agente*, puede ser substituida por (2). La conclusión es que «Algo acaece... como...» es demasiado general como para garantizar la topicidad neutral, que (1) y (2) no dicen realmente lo mismo, y que cualquier maniobra tendente a eliminar la dificultad incluirá un rasgo tópico asociado a la experiencia del agente[9].

La neutralidad tópica plantea a la TI una dificultad importante. Es uno de los factores que hará que algunos de sus defensores prominentes (Smart incluido) se conviertan al eliminativismo.

2. *Las propiedades fenoménicas*

Las identidades que propone la TII deben ser entendidas como identidades estrictas (*item* B de la matriz teórica). Pero ¿qué debe entenderse por

9. Adaptación de un argumento desarrollado por Cornman (1962).

«identidad estricta» en tal contexto? Los teóricos de la identidad no respondieron a esta pregunta de una manera explícita. Excluyeron los casos de identidad por contigüidad espacial o temporal, y sugirieron que *A* y *B*, digamos, son estrictamente idénticos si «A» y «B» denotan lo mismo. La respuesta dista de ser satisfactoria. Los críticos de la TII llenaron prontamente el vacío. Interpretaron que la única respuesta adecuada a la pregunta por la identidad estricta es en términos del Principio de Indiscernibilidad de los Idénticos: *A* y *B*, dos objetos cualesquiera, son idénticos si comparten todas y cada una de sus propiedades. A continuación sumaron una objeción tras otra. *El Dolor* tiene propiedades que nunca puede tener *Disparos de las fibras-c*, y viceversa (para seguir con el socorrido ejemplo). No apelamos a las mismas vías cognoscitivas para *El Dolor* que para *Disparos de las fibras-c*, etc. Si tal es el caso, es obvio que no puede haber identidad entre los fenómenos mentales y los fenómenos físicos [10].

La respuesta a este tipo de crítica comienza con una operación de descarte. Las objeciones que apelan a modalidades (cognoscitivas o de otro tipo) son desechadas porque no se adecuan a los cánones del Principio de Indiscernibilidad de los Idénticos. Respecto de las objeciones restantes (es decir, de las que apelan a aspectos fenómenicos) se propone la estrategia siguiente. Se reitera el primer paso de la respuesta al problema de la neutralidad tópica, esto es, se rechaza la idea de que los fenómenos mentales sean «objetos» (*El dolor, La sensación*) con propiedades específicas, con los que los agentes están relacionados. Los fenómenos mentales son concebidos, entonces, como eventos/estados/procesos que resultan ser propiedades de los agentes. De esta manera, se abandona la referencia a *El Dolor*, y similares, y se pasa a *Tener experiencia de dolor* o a *Tener dolor*. En un segundo paso, se niega que en un sentido estricto existan los *Dolores*, las *Sensaciones*, las *Postimágenes*, etc. Lo que existe son unos ciertos cambios en los agentes que, digamos, *tienen una experiencia de dolor*, o *una sensación de verde* o *una postimagen amarillenta*. Finalmente, se señala que las identidades valen entre eventos/estados/procesos mentales y eventos/estados/procesos neurofisiológicos. Las identidades que propone la TII son, así, ontológicamente homogéneas. A la objeción adicional de que los contenidos cualitativos de las experiencias quedan sin reducir, se responde que son aspectos físicos esenciales de ciertas propiedades físicas.

La estrategia es buena. Bloquea, en principio, el argumento de que, por ejemplo, *El Dolor* tiene propiedades (ser agudo, ser intermitente) que no tiene *Disparos de las fibras-c*. El defensor de la TII puede ahora responder a sus críticos que los eventos mentales no tienen propiedades de

10. Borst (1980) incluye parte de la polémica. Carruthers (1991) analiza una serie de argumentos que adoptan esa estrategia.

ese tipo, que ellos mismos son propiedades complejas de los agentes, y que en un cierto sentido el aspecto fenoménico se torna una parte del fenómeno mismo.

Las conclusiones a las que conduce esta estrategia conceptual han originado una polémica en la que ronda permanentemente el fantasma del dualismo (por cierto, no sólo el del dualismo substancialista). Lo que se cuestiona, básicamente, es que esa estrategia (y para muchos críticos, cualquier otra) permita explicar el *status* ontológico y gnoseológico de los contenidos cualitativos de ciertos eventos mentales y de los estados de conciencia que los acompañan. Los críticos sostienen que el contenido cualitativo de ciertos estados mentales, el *quale*, no es (no puede ser) lo propio de ningún estado físico. Hay situaciones reales y situaciones contrafácticas que lo prueban. Una persona provista de todo el conocimiento imaginable acerca de sus estados físicos, no tiene (no puede tener) conocimiento de los tipos de *qualia* que no ha experimentado. Argumentos de este tipo parecen implicar que hay un *plus* ontológico que no encaja (que no puede encajar) en el marco fisicalista de la TII. Los *qualia* son elementos intrínsecos de las experiencias. En consecuencia, el intento de reducirlos a lo físico está condenado al fracaso. La propuesta de concebirlos como aspectos físicos de propiedades físicas tiene, además, la consecuencia absurda de que los fenómenos mentales a los que nos referimos de manera corriente, dejan de ser lo que son. Por otra parte, sostener que los contenidos cualitativos son, en un sentido literal, aspectos físicos esenciales de ciertas propiedades físicas, implica introducir ingredientes mentales en el mundo físico. Y ello resulta inaceptable.

No es del caso desarrollar aquí los detalles de esta intrincada discusión [11], pero corresponde apuntar que los problemas que plantean los *qualia* y los estados de conciencia, no afectan únicamente a la TII. Su impacto es más amplio. Toda teoría de los fenómenos mentales insuflada de aires fisicalistas tiene ante sí la difícil tarea de «explicar» los *qualia* y los estados de conciencia que les corresponden. No es casual que las expresiones más recientes de la polémica involucren al funcionalismo. Por otra parte, el problema de los *qualia* y de los estados de conciencia es a las posiciones fisicalistas lo que el problema de la interacción es al dualismo substancialista: un tema teóricamente urticante, una especie de talón de Aquiles. Soy consciente de que este no es un argumento filosófico. Sólo es un intento de defender a la TII apelando a aquello de «mal de muchos». Pero creo que, de algún modo, la defensa posee cierta miga teórica. La mención de las dificultades que también afectan al funcionalismo y al dualismo substancialista es pertinente a la hora de evaluar los méritos de la

11. Véase Stevenson (1960) y Baier (1962). Para los planteos más recientes, véase los trabajos compilados en Lycan (1990) y la bibliografía correspondiente. Ver también *infra* el capítulo de García Suárez, pp. 353-383.

TII. Respecto de los *qualia*, la TII tiene las mismas dificultades que afectan al funcionalismo. Y respecto del dualismo substancialista, la versión física de los *qualia* que proporciona la TII es más creíble, menos *ad hoc*, que la historia interaccionista, paralelista o epifenomenista.

3. *Las identidades necesarias*

Una de las tesis centrales de la TII es que las identidades que postula expresan verdades contingentes (*items* C, D y E de la matriz teórica). «El dolor es disparos de las fibras-c» es equiparado a enunciados como «Los genes son segmentos de moléculas de ADN» o «El calor es energía cinética media», que son casos típicos de enunciados sintéticos y *a posteriori*. El carácter contingente de las identidades tiene una importancia especial. La TII se presenta como una teoría filosófica que hipotetiza acerca del avance del conocimiento científico y de la posibilidad de formular identidades de fenómenos mentales con estados neurofisiológicos. Esas identidades heredan, pues, las características corrientes de los enunciados científicos. Que las identidades sean contingentes implica que pueden ser falseadas, cambiadas (según evolucione la ciencia) y aun abandonadas *in toto* (si se llegara a constatar la imposibilidad global de la empresa). El carácter contingente de la relación entre los fenómenos mentales y los estados neurológicos implica, además, la *posibilidad* de que fenómenos que son reconocidos como mentales no tengan una contrapartida neurofisiológica, y de que ciertos estados neurofisiológicos no tengan una contrapartida mental. Esta es una consecuencia más que curiosa: por razones distintas, la TII coincide en este punto con el dualismo cartesiano.

Los defensores de la TII no fundan el carácter contingente de las identidades en un argumento explícito. Ese carácter surge por añadidura del marco científico en el que las insertan. Presuponen, sin embargo, una teoría semántico-ontológica específica. Se trata de la teoría empirista de la referencia, la necesidad lógica y las propiedades esenciales. Sus tesis básicas son conocidas: 1. la necesidad emana del significado de las palabras y de las convenciones lingüísticas (toda necesidad es *de dicto*); 2. el significado de los términos generales y, por ende, el significado de los términos de clases naturales, surge de criterios convencionales; 3. los problemas acerca de la clasificación de las entidades son, en consecuencia, problemas que involucran a las convenciones lingüísticas vigentes; 4. en los casos conflictivos no es posible producir una refutación empírica, sino sólo proponer el cambio de convenciones, es decir, el cambio de significado; 5. las propiedades esenciales de las cosas dependen de las descripciones que les damos; 6. el carácter contingente de un enunciado depende de que sea refutable; 7. necesidad y carácter *a priori* coinciden.

La aplicación estricta de esta teoría semántico-ontológica tiene, en

principio, implicaciones indeseables para las identidades que postula la TII. Consideremos nuevamente «El dolor es disparos de fibras-c». «El dolor» y «disparos de fibras-c» son términos generales con distinto significado. Según la teoría empirista esto equivale a afirmar que los términos están regulados por convenciones diferentes que especifican en el mundo clases naturales diferentes. De esto se sigue que «El dolor es disparos de fibras-c» nunca puede ser verdadero [12].

La situación de la TII se torna más comprometida aún cuando se la incluye en un contexto semántico-ontológico distinto al de la teoría empirista de la necesidad y las clases naturales. Kripke ha argumentado que 1. «el dolor» y «disparos de las fibras-c» son designadores rígidos, es decir, términos que se refieren a lo mismo en todo mundo posible en el que tienen referencia; 2. *El Dolor* y *Disparos de fibras-c* tienen propiedades esenciales que, en tanto tales, son independientes de criterios de carácter convencional; 3. es esencial a *El Dolor* ser una experiencia y a *Disparo de fibras-c* ser un arreglo molecular; 4. si «El dolor es disparo de las fibras-c» es verdadera, entonces es necesariamente verdadera [13]. La conclusión es que el defensor de la TII tiene que ofrecer un argumento filosófico serio que fundamente la intuición de la contingencia y la posibilidad de que haya fenómenos mentales sin contrapartida neurofisiológica y estados neurofisiológicos sin contrapartida mental. Kripke considera que tal argumento no es posible. Por supuesto que el defensor de la TII no puede aceptar que sus identidades expresen verdades necesarias. Kripke concluye que su planteo no sólo vale contra la TII, sino contra la tesis general de que los estados mentales son estados de naturaleza física [14].

Estos argumentos introducen en la discusión de la TII y, en general, de toda posición de raigambre fisicalista, la necesidad de aclarar los supuestos semántico-ontológicos que la sustentan. El tema es legítimo e importante. Por cierto que hay defensas posibles para la TII y el fisicalismo [15].

4. *La realizabilidad variable*

Una manera de caracterizar con estrictez la tesis básica de la TII (*items* A y B de la matriz teórica), es ésta:

(a) Para cada tipo psicológico *P* (*Tener dolor*, digamos) hay un único tipo físico *F* (*Tener disparadas las fibras-c*, digamos), tal que *P* es coextensivo con *F*. La coextensividad tiene un carácter nomológico.

12. Sigo en estos puntos la argumentación de Boyd (1980).

13. Kripke (1971). Los puntos 1-4 pretenden sintetizar algunas de las tesis básicas que formula Kripke. Véase la obra citada y Kripke (1972).

14. Los argumentos de Kripke han dado lugar a una extensa bibliografía. Véase, entre otros, Feldman (1974), Boyd (1980), Felmand (1980), MacDonald (1989) y la bibliografía que citan.

15. Véanse los trabajos citados de Feldman y Boyd, por ejemplo. Ver también Valdés (1980)

(b) Que *P* y *F* sean coextensivos significa que, dado un sistema S al que se atribuyen estados psicológicos, *S* instancia *P* en un tiempo *t* si y sólo si *S* instancia *F* en *t*.

(c) *F* no varía de especie a especie ni con los tipos de constitución estructural.

Esta caracterización expone con claridad un rasgo típico de la TII. Las identidades que postula se dan entre *tipos* o *propiedades* mentales (psicológicas) y *tipos* o *propiedades* neurofisiológicas. Por supuesto que sostener la identidad de tipos no excluye sostener la identidad de los *casos* en los que se realizan o instancian (por ejemplo, *tener dolor por parte de A en t* es *tener disparadas las fibras-c por parte de A en t*), aunque la conversa no se da necesariamente. (La cursiva en minúscula refiere a los casos de los fenómenos tipo referidos con cursivas en mayúscula; recuérdese la convención introducida en II.B). La caracterización expone, además, otro rasgo típico de la TII: la base neurofisiológica que postula es única, en el sentido de que no varía de especie a especie, ni con la constitución físico-química. La constitución estructural supuesta es la del sistema nervioso humano.

Todo esto tiene la consecuencia de que cada estado mental es siempre y ecuménicamente un único tipo de estado neurofisiológico. Por lo tanto, la atribución de estados psicológicos a un sistema cualquiera (actual o contrafáctico), supone que ese sistema posee la base neurofisiológica requerida. Dada esa base, corresponde la atribución de estados psicológicos. La duda que surge es si esta consecuencia está justificada; si no es demasiado restrictiva. Muchos filósofos piensan que lo es y que, en consecuencia, carece de un fundamento razonable.

En unos párrafos célebres, Putnam argumentó que respecto de *El dolor*, por ejemplo, el defensor de la TII tiene que especificar un estado físico-químico de cualquier organismo O tal que O tiene dolor si y sólo si O tiene un cerebro con una estructura físico-química adecuada y el cerebro de O está en ese estado. Al mismo tiempo, no tiene que ser un estado posible del cerebro de ningún O que no pueda sentir dolor. Supongamos que podemos descubrir tal estado, entonces será también el estado del cerebro de un ser extraterrestre, por ejemplo, con prescindencia de que hayamos llegado a suponer, siquiera, que lo que ese ser tiene *sea* dolor. Por otra parte, como la TII sostiene que *todo* estado psicológico es un estado cerebral, su refutación resulta ser extremadamente fácil. Basta con encontrar un tipo o propiedad psicológica que valga respecto de dos especies distintas pero cuyos correlatos neurofisiológicos sean diferentes. Es altamente probable que esa situación se dé [16].

El argumento puede plantearse respecto del propio cerebro humano.

16. Putnam (1967), a quien se debe la caracterización estricta.

El cerebro tiene una plasticidad tal que la identificación de tipos psicológicos con tipos neurofisiológicos resulta prácticamente imposible. Respecto de una misma persona, cabe dudar que el evento neurofisiológico en que consiste su sentir dolor en un momento dado, sea el mismo en el que consiste su sentir dolor en otro momento[17]. La TII resulta ser notoriamente restrictiva e irreal. Ni los neurofisiólogos más optimistas estarían dispuestos a avalarla como programa científico.

La crítica está íntimamente relacionada con la prédica funcionalista. *Tener dolor* no es idéntico a *Tener disparadas las fibras-c*. *Tener dolor* se identifica con la propiedad de un estado que juega un determinado rol causal o funcional. Ese rol es independiente de la base física en la que se implementa. Las peculiaridades de la base física no afectan la psicología de los organismos[18].

Volveré sobre el tema en la Sección V.

IV. EL ELIMINATIVISMO

Señalé al comienzo que el filósofo que se dispone a lidiar con el problema mente-cuerpo tiene que tomar en cuenta ciertas convicciones de sentido común. La TII da por supuestas esas convicciones, en particular la distinción entre fenómenos mentales y fenómenos físicos. Más aún, la TII reconoce la *realidad* de los fenómenos mentales y, consiguientemente, los considera una parte legítima del moblaje del mundo. Sus propiedades son, en lo esencial, las que se les atribuye de manera corriente. El defensor de la TII cree que esas propiedades, o al menos una parte pertinente de ellas, pueden ser acomodadas dentro de su esquema teórico. Pero algunas de las dificultades más serias que afectan a la TII surgen de ese reconocimiento. Tal es el caso de los problemas planteados en III.1 y III.2.

Adviértase que las soluciones propuestas a cada uno de ellos implican, de alguna manera, reformulaciones de las convicciones básicas. La forma normal que propone Smart para lograr reportes tópicamente neutrales y la sugerencia de dejar a un lado la referencia a *El dolor* para hablar de *Tener experiencias de dolor* son, cada una a su manera, reformulaciones de ese tipo. Pero las dificultades subsisten. Para superarlas puede recurrirse a una solución extrema: el eliminativismo (también llamado «Materialismo eliminativista»).

La matriz teórica del eliminativismo[19] es la siguiente:

17. La expresión «su sentir dolor», referida a la persona, y la expresión asociada «el sentir dolor por parte de A (un agente cualquiera)», gramaticalmente molestas, hacen referencia a eventos.
18. Sobre funcionalismo, ver el capítulo 2 de este volumen (García Carpintero), pp. 43-76.
19. Feyerabend (1963a).

(A) Los defensores de la TII se equivocan cuando defienden la hipótesis empírica:

(H) X es un proceso mental de tipo *A* si y sólo si *X* es un proceso de tipo *a* del sistema nervioso central.

H implica que los procesos mentales tienen rasgos físicos, pero también implica que algunos procesos físicos tienen rasgos mentales. La TII se compromete, así, con una posición dualista (un dualismo de propiedades) que le resulta imposible superar.

(B) La polémica entre dualistas y fisicalistas no se puede decidir en base a la discusión acerca del *status* empírico de H. De hecho, H es falsa.

(C) La manera corriente de hablar de los fenómenos psicológicos y los fenómenos físicos involucra una teoría acerca de las personas. Los méritos eventuales de una teoría neurofisiológica deben ser apreciados con independencia de ese marco teórico. En general, las teorías científicas se desarrollan sin presuponer marcos teóricos que les sirvan de fondo.

(D) Los fenómenos psicológicos de los que hablamos corrientemente, no existen. Si por alguna razón deseamos seguir hablando de ellos en todo o en parte, el procedimiento correcto consiste en definirlos a partir de una teoría científica, no a la inversa.

(E) Se sigue de lo anterior que las propiedades que se suelen denominar «psicológicas», no son (no pueden ser) coextensivas con las propiedades físicas.

(F) H pretende ser, típicamente, una ley puente. La no existencia de leyes puente no constituye, al menos en este caso, un criterio suficiente para medir el éxito de la teorización científica.

(G) Los estados neurofisiológicos son estados internos de los agentes; el avance de la neurociencia brindará el conocimiento que se precisa para describir y explicar las capacidades y los procesos cognitivos.

Los eliminativistas complementan su planteo con el análisis de ejemplos que extraen de la historia de la ciencia. Esos ejemplos muestran cómo el avance de las teorías científicas ha llevado a abandonar maneras corrientes de hablar y/o de teorizar. La teoría fisiológica de la epilepsia no se ve perjudicada por haberse abandonado la teoría acerca de «la enfermedad divina». La identificación de las bacterias y de los virus como la causa de ciertas enfermedades permite descartar la apelación a los demonios que hace el hechicero de la tribu. Las brujas «existieron» en la Edad Media, pero las teorías actuales sobre ciertos tipos de histeria femenina permiten sostener que no hubo brujas. En general, podemos afirmar la no existencia de una entidad cuando descubrimos una manera novedosa de explicar un fenómeno que previamente se explicó apelando a dicha entidad, y esa explicación novedosa nos permite dar cuenta de los enunciados observacionales pertinentes[20].

20. Rorty (1965).

Los eliminativistas realzan la capacidad descriptiva del lenguaje y llaman la atención sobre los mecanismos que la constituyen y los criterios que empleamos para evaluarla. Al sostener que los fenómenos mentales existen realmente, los no eliminativistas (incluidos los partidarios de la TII) suponen que los fenómenos mentales proporcionan una mejor descripción del mundo. Esto es, precisamente, lo que el eliminativista cuestiona. La descripción más completa y efectiva del mundo apela a los fenómenos físicos y proviene del desarrollo de la ciencia. Una descripción completa tal excluye el lenguaje con contenido psicológico.

Una discusión detallada del eliminativismo supone sopesar el valor de los ejemplos históricos que utiliza, la viabilidad de la concepción de las teorías científicas que presupone, la prognosis que hace del desarrollo de la ciencia y los méritos de una descripción fisicalista completa del mundo. Todos ellos son temas dignos de reflexión. Pero el carácter extremo del eliminativismo genera, por lo general, una actitud distinta. Se argumenta que aceptar los fenómenos mentales como entidades legítimas del mundo forma parte esencial de la TII, que la atracción que posee la TII emana, en gran medida, de que intenta dar cuenta de ellos dentro de una matriz teórica austera, y que adoptar el eliminativismo es reconocer, en definitiva, el fracaso del programa. Estas consideraciones, sin duda aceptables, no resuelven por sí mismas el problema. Los eliminativistas consideran que el programa de la TII es inviable, tal como se lo pergeñó originariamente. La carga de la prueba corresponde al defensor de tal programa.

Corresponde aclarar que el eliminativismo descrito es el «eliminativismo clásico», distinto del llamado «eliminativismo reciente» [21]. Esta nueva versión del eliminativismo ha florecido a la vera del funcionalismo y ha generado una importante polémica en torno al *status* de la psicología de sentido común (*folk psychology*) y de los compromisos realistas a asumir respecto de los estados y procesos mentales [22]. Es bueno recordar en este punto las observaciones que formulé al final de III.3 acerca de la evaluación de la TII. Los problemas relacionados con el fisicalismo parecen superar los méritos de las propuestas teóricas. Tal como ocurre con el problema de los *qualia* y de la conciencia, la tentación del eliminativismo afecta por igual a la TII y al funcionalismo.

V. LA TEORIA DE LA IDENTIDAD DE ROL CAUSAL

La teoría de la identidad de rol causal (también llamada materialismo funcionalista, teoría causal de la mente, funcionalismo causal, funciona-

21. La terminología es de Lycan(1990). Ver pp. 245-273 de esta compilación.

22. Véase Churchland (1981), Stich (1983) y Fodor (1985) (1987). Sobre psicología de sentido común, véase Greenwood (1991).

lismo analítico; TIRC, en lo sucesivo) no sólo significa un avance teórico evidente *vis-a-vis* la TII, sino que es una contribución substancial al problema mente-cuerpo. Además, la TIRC posee una elegancia teórica poco común.

Hay conceptos que exhiben, típicamente, rasgos causales. Un ejemplo trivial es el concepto de veneno. Un veneno es algo que cuando es absorbido por un determinado tipo de organismo, tiene la aptitud de causar su muerte o, al menos, su deterioro. Un veneno *actúa* de una determinada manera, pero para producir efecto tiene que ser asimilado en una dosis adecuada y no tienen que existir factores que lo neutralicen. En suma, un veneno es lo que es capaz de producir un cierto tipo de efectos, esto es, de cumplir un cierto rol causal. Queda a la ciencia determinar cuáles son las substancias venenosas, es decir, cuáles son las substancias que pueden llegar a ocupar, en los casos específicos, el rol causal de veneno.

La distinción es importante. Una cosa es especificar el rol causal de un cierto concepto; otra cosa es especificar qué o quién es el ocupante de ese rol. Cuando la distinción se aplica a los conceptos mentales, conduce a lo siguiente: el concepto de un estado mental tipo *x*, es el concepto de algo que de un modo característico causa determinados efectos y es a su vez efecto de ciertas causas características. Los efectos de *x* son patrones de conducta de la persona que tiene *x*. Las causas de *x* son eventos en el entorno de esa persona. Una descripción del rol causal de *x* incluye la descripción de sus causas y efectos típicos, así como las conexiones causales posibles de *x* con otros estados mentales. Adviértase que lo que se especifica es el rol causal que *x tiene* en su condición de mediador interno entre las causas del entorno y los efectos conductuales. No se afirma que *x sea* ese rol causal. Esto distingue a la TIRC del conductismo.

El desarrollo de este planteo involucra dos fases. La primera, conceptual, ofrece análisis detallados de los diferentes conceptos mentales tipo, explicitando así su significado. Se trata de una investigación *a priori* que culmina con la formulación de enunciados analíticamente verdaderos. La segunda, empírica, ubica y describe los estados físico-químicos tipo del cerebro que ocupan los roles causales atribuidos a los conceptos mentales [23].

La primera tarea es filosófica, la segunda científica. El planteo culmina con el establecimiento de las identidades contingentes que se dan entre los roles causales y sus ocupantes.

La versión elaborada de la TIRC introduce cambios de importancia en este esquema inicial. Su matriz teórica es la siguiente [24]:

23. Armstrong (1968).
24. Lewis (1966) (1969) (1970) (1972) (1974) (1980).

(A) La hipótesis general es la de la TI: toda experiencia mental tipo *es* (idéntica a) algún estado físico (neurológico) tipo.

(B) La TII presupone que 1. los avances científicos hacen posible la formulación de leyes puente que identifican algunas entidades de una teoría con las de otra teoría; 2. las identidades interteóricas («Agua es H_2O») no se descubren, sino que se construyen, y 3. lo que justifica a las identidades teóricas es la economía ontológica que implican. Estas presuposiciones son erróneas. Las identidades psicofísicas son implicadas por las teorías que las hacen posibles. Una teoría, la psicología de sentido común, permite la introducción de términos caracterizados por su rol causal. Otra teoría, la neurofisiología, en conjunción con la primera, implica las identidades psicofísicas. El significado de los términos pertinentes y la neurofisiología conducen, necesariamente, a las identidades psicofísicas.

(C) Lo anterior permite formular el esquema argumentativo básico:

> El estado mental *M* = El ocupante del rol causal *R* (por definición de *M*).
> El estado neural *N* = El ocupante del rol causal *R* (por la teoría neurofisiológica).
> En consecuencia, El estado mental *M* = El estado neural.

(D) Las adscripciones de experiencia tienen la misma denotación que las adscripciones de estados neurofisiológicos, pero poseen distinto sentido. Las primeras se refieren a un estado mediante la especificación de su rol causal. Las segundas se refieren a él mediante descripciones detalladas.

E) La neutralidad tópica se logra mediante un procedimiento que permite eliminar los términos mentales provenientes del marco teórico de la psicología de sentido común (*T*). Si se ponen en conjunción los truismos de *T* y se identifican los términos teóricos (términos *T*) que corresponden a los estados mentales, se obtiene el postulado de *T*. Esos términos teóricos funcionan como términos singulares (*S1...Sn*). Los demás términos que se introducen son los términos *O*. El postulado de *T* dice que los términos *T* ocupan ciertos roles causales y que tienen relaciones causales entre sí y con las entidades nombradas por los términos *O*. En un segundo paso, se suplantan los términos *T* por variables. Se prefijan entonces cuantificadores existenciales y se obtiene la Oración Ramsey de *T*, que dice que *T* tiene al menos una realización. La Oración Ramsey Modificada dice que *T* tiene una única realización.

(F) Los conceptos y nombres corrientes de los estados mentales son no rígidos. A qué estado se aplica un concepto y la palabra correspondiente, es una cuestión contingente. El concepto de dolor y la palabra «dolor», por ejemplo, se aplica en nuestro mundo a un cierto estado neu-

ral, pero no en otro mundo. Un cierto estado ocupa un rol causal *para una población*. Toda vez que un miembro de esa población está en ese estado, está en el estado que tiene el tipo de causas y de efectos dados por el rol. La denotación del término varía, entonces, de población en población, pero el concepto expresado por el término correspondiente es fijo.

(G) Las experiencias, en tanto procesos o actividades introspectibles, son estados físicos. Pero debe distinguirse la experiencia en sí del atributo que se predica de quien tiene la experiencia. La primera corresponde al estado que ocupa un cierto rol causal, la segunda es el atributo de estar en el estado, cualquiera sea él, que ocupa ese rol causal. Esta distinción permite hacer frente al argumento acerca de la no sinonimia de las adscripciones de estados mentales y las adscripciones de estados neurales.

La TIRC es la oferta teórica más elaborada y más convincente de la Tesis de la Identidad. Pero, como es de prever, ha sido objeto de una serie de críticas[25]. Algunas apuntan a su apego al marco teórico de la psicología de sentido común; otras a su insistencia en las identidades tipo/tipo; otras a las dificultades que afectan, en general, a los enfoques causales y funcionalistas. No puedo entrar aquí en un análisis detallado. Dedicaré el resto de este capítulo a las críticas a la TIRC y, en general, a la TI, basadas en el argumento de la realizabilidad variable y sus consecuencias antirreduccionistas.

VI LA REALIZABILIDAD VARIABLE Y EL REDUCCIONISMO

El argumento de la realizabilidad variable (III.4) ha jugado un papel importante en el descrédito de la TI y, consiguientemente, en la entronización del funcionalismo. El argumento niega que sea posible identificar los estados psicológicos con estados neurológicos, porque su relación no es de uno a uno sino de uno a muchos. El punto es que no existe, ni puede existir, una clase natural física única que pueda correlacionarse con cada clase natural genérica de la psicología, de la manera que la TI exige. Las propiedades psicológicas se realizan (instancian, implementan, ejemplifican) en bases físicas heterogéneas. En consecuencia, no pueden ser identificadas (ni son identificables) con una de ellas ni con la disyunción de todas ellas. *Realizabilidad* es una relación distinta a *identidad*. Y no sólo por razones formales. Su papel metodológico es diferente. La identidad cuadra en el marco de la concepción aceptada de la reducción

25. Nagel (1970), Block (1978), Shoemaker (1981), MacDonald (1989), entre otros.

y de la unidad de la ciencia. La realizabilidad requiere una concepción distinta[26].

Putnam dio al argumento un fundamento fáctico: la imposibilidad de correlacionar estados psicológicos tipo con estados neurológicos tipo se debe a circunstancias empíricas (ver detalles en III.4). Pero, a menudo, el argumento se plantea como una tesis conceptual: que una propiedad psicológica se pueda realizar en un conjunto heterogéneo de propiedades físicas, es incongruente. Las propiedades psicológicas, concebidas como propiedades causal/funcionales, son propiedades de segundo orden (es decir, son propiedades de las propiedades de los estados mentales), y la especificación de tal tipo de propiedades no impone restricciones en cuanto a su implementación física. La versión empírica afecta a la TI, en cualquiera de sus versiones. La versión conceptual dispara directamente contra la TII y, por elevación, contra la TIRC; esta introdujo el análisis causal/funcional de las propiedades psicológicas dentro del ámbito de la TI y, según el argumento, no es fiel a las implicaciones de tal análisis.

Se suele atribuir al argumento un significado filosófico peculiar. Si es imposible establecer identidades entre los tipos psicológicos y los tipos neurofisiológicos, la *reducción* de lo mental a lo físico es imposible. Esto desbarata al reduccionismo como tesis filosófica, clava una pica en el corazón mismo de la TI (*item* G de la matriz teórica de la TII e *items* B, C y E de la matriz teórica de la TIRC) y abre las puertas a los programas antirreduccionistas (funcionalismo, anomalismo). Pero adviértase que la imposibilidad de establecer identidades entre tipos, no afecta la posibilidad de establecer identidades entre casos de estados/eventos mentales y casos de estados/eventos neurofisiológicos. Sostener la viabilidad de la identidad de tipos implica sostener la de la identidad de casos; pero la conversa no se da (2.4). Al partidario de la realizabilidad variable le es permitido, entonces, admitir que hay identidades como el *tener dolor* por parte de *A* en un tiempo *t* ES el *tener disparadas la fibras-c* por parte de *A* en el tiempo *t*. Es decir, le es permitido adherir al Fisicalismo de Casos: todos los *eventos* de los que hablan las ciencias son eventos físicos[27].

El cuadro crítico que genera el argumento va tomando forma: si al ingrediente de base que proporciona el fisicalismo se le agregan cantidades adecuadas de funcionalismo, antirreduccionismo e identidad de casos, se obtendrá el condimento que sazona la dieta teórica actual de muchos filósofos de la mente.

Pero ¿tiene el argumento de la realizabilidad variable las consecuencias dramáticas que se le atribuye *vis-a-vis* la TI? Mi impresión es que al-

26. Para la caracterización de la reducción interteórica véase el clásico capítulo 11 de Nagel (1961). Para la polémica sobre la unidad de la ciencia y el status de las «ciencias especiales», véase Oppenheim y Putnam (1958) y Fodor (1974).

27. Fodor (1974).

gunas pueden manejarse razonablemente bien dentro de la TI; otras consecuencias son, en realidad, tesis en favor de esquemas teóricos alternativos que se introducen bajo la forma de un argumento general. Y esto requiere una estrategia discursiva diferente. He aquí algunos puntos mínimos respecto de esta evaluación.

La versión empírica del argumento puede ser neutralizada con bastante éxito. Del hecho de que dos sistemas nerviosos estén compuestos y estructurados de manera diferente, no se sigue que no puedan estar en un mismo estado físico. La razón es simple. Las diferencias que se detectan en las bases físicas de dos sistemas nerviosos, siempre son relativas a un cierto esquema clasificatorio. Nada impide que esos sistemas estén en un mismo estado con respecto a otro sistema clasificatorio. Esto es típico de la investigación científica y de los procedimientos clasificatorio/explicativos que involucra[28]. Pero concedamos que los correlatos físicos de los estados psicológicos tipo dependen de las peculiaridades de las especies y aun de los individuos. ¿Se siguen de este reconocimiento las consecuencias del argumento? Pareciera que no. Es posible especificar en los correlatos físicos rasgos comunes que permitan formular caracterizaciones *relativas a la especie*. En consecuencia, es posible formular generalizaciones del tipo «Los miembros de la especie *E* tienen *x* (un estado psicológico tipo) cuando están en el estado cerebral *c*». Y si cada uno de los tipos psicológicos tiene una realización física en una especie, es posible *reducir localmente* esos tipos a la teoría física de la especie. Las leyes puente necesarias para la reducción serán entonces *leyes puente locales*. El caso de las peculiaridades individuales puede manejarse apuntando a su irrelevancia (dentro de un marco teórico general) o bien al carácter probabilístico de las leyes propias de la psicología[29].

Se puede objetar que todo esto implica que no vamos a poder contar con un concepto general para cada tipo mental; que tendremos, por ejemplo, dolor (para los humanos), dolor (para los moluscos), dolor (para los *X*), etc. Esta objeción, sin duda importante, puede ser respondida a través de dos estrategias distintas: 1. no hay razones filosóficas o cognoscitivas de peso que hagan suponer que la TI se va a ver afectada si «sólo» se logran identidades relativas a cada especie[30]; 2. contamos con un concepto general de dolor, que es el mismo para organismos diferentes. Ello es, precisamente, lo que permite determinar las variaciones que se dan en la denotación de «dolor» de especie a especie[31].

Cuando se da al argumento un giro conceptual, la discusión se torna mucho más complicada. Como he señalado, la versión conceptual del ar-

28. Kim (1972).
29. Kim (1982) (1989) (1992).
30. Kim (1972). Véase un desarrollo detallado de este punto en Kim (1992).
31. Lewis (1969) e *item* F de la matriz teórica de la TIRC.

gumento es una pieza clave en la estrategia antirreduccionista. Discutirlo involucra, pues, meter mano en un conjunto heterogéneo de temas que van de la tesis de la unidad (o la desunión) de la ciencia y la eventual autonomía de la psicología, al *status* de cierto tipo de predicados (en particular, los predicados disyuntivos), la causación mental y ciertas implicaciones ontológicas y metodológicas. Es imposible intentar, siquiera, una síntesis de esta discusión. Lo crucial es advertir que todos esos temas suponen un cierto tipo de decisiones que están más allá o, a menudo, más acá del argumento de la realizabilidad variable en sí mismo[32].

La tesis de la identidad se propuso lidiar con el dualismo de estirpe cartesiana y superar, al mismo tiempo, la «solución» conductista. Ese aspecto de su programa teórico tuvo un éxito razonable. Pero los méritos de la tesis de la identidad superan, obviamente, ese objetivo contextual. Es difícil entender la situación actual de la filosofía de la mente sin tener una comprensión cabal de la TI, de sus variantes y de sus dificultades. Si el éxito de una teoría filosófica se mide por el impacto de las matrices teóricas que genera, las elaboraciones conceptuales que incita y las controversias que produce, la TI fue y es, fuera de toda duda, una teoría filosófica importante. Por ello merece un estudio serio. Pero su valor no es meramente histórico. En estos tiempos de realizabilidad, superveniencia, funcionalismo, teleología y emergentismo, es bueno tenerla presente. A decir verdad, explicar cómo se relacionan las propiedades psicológicas con las propiedades físicas, tratando de ser fiel a las presuposiciones del fisicalismo, sigue siendo un misterio.

BIBLIOGRAFÍA

Armstrong, D. M. (1965), «The Nature of Mind»; incluido en Borst, 1970.
Armstrong, D. M. (1968), *A Materialist Theory of the Mind*, Routledge, London.
Armstrong, D. M. (1981), *The Nature of Mind and Other Essays*, Cornell University Press, Cornell.
Baier, K. (1962), «Smart on Sensations»: *Australasian Journal of Philosophy*, *40*; incluido en Borst, 1970.
Bieri, P. (1981), «Materialismus. Einleitung», en P. Bieri (comp.), *Analytische Philosophie des Geistes*, Hain, Königstein/Ts.
Block, N. (1978), «Troubles with Functionalism», en C. W. Savage (comp.), *Minnesota Studies in the Philosophy of Science IX*, University of Minnesota Press, Minneapolis; incluido en Block, 1980 y Rabossi, s/f.
Block, N. (comp.) (1980), *Readings in Philosophy of Psychology*, Harvard University Press, Cambridge, Mass.

32. Véase Kim (1992). En el capítulo 7 de este volumen se desarrolla y discute este importante núcleo temático.

Borst, C. V. (comp.) (1970), *The Mind-Brain Identity Theory*, MacMillan, London.
Boyd, R. (1980), «Materialism Without Reductionism: What Physicalism Does Not Entail», en Block, 1980.
Bunge, M. (1980) *The Mind-Body Problem. A Psychobiological Approach*, Pergamon Press, Oxford. V.e.: B. García Noriega, Tecnos, Madrid, 1988.
Campbell, K. (1970), *Body and Mind*, University of Notre Dame Press, Notre Dame, [2]1894.
Carruthers, R. (1991), *Introducing Persons: Theories and Arguments in the Philosophy of Mind*, Routledge, London.
Churchland, P. M. (1981), «Eliminative Materialism and Propositional Attitudes»: *Journal of Philosophy, 78*; incluido en Rabossi, s/f.
Churchland, P. M. (1988), *Matter and Consciousness*, MIT Press, Cambridge, Mass. V.e.: M. Mizraji, GEDISA, Barcelona, 1992.
Corruthers W. C. (1962), «The Identity of Mind and Body»: *Journal of Philosophy, 59*; incluido en Borst, 1970 y Rosenthhal, 1971.
Esquivel, J. (comp.) (1982), *La polémica del materialismo*, Tecnos, Madrid.
Feigl, H. (1958), «The "Mental" and the "Physical"», en H. Feigl, M. Scriven y G. Maxwell (comp.), *Minnesota Studies in the Philosophy of Science* II, University of Minnesota Press, Minneapolis.
Feigl, H. (1960), «Mind-body, not a Pseudo-Problem», en Hook, 1960.
Feigl, H. (1967), *The «Mental» and the «Physical». The Essay and a Postscript*, University of Minnesota Press, Minneapolis.
Feldman, F. (1974), «Kripke on the Identity Theory»: *Journal of Philosophy, 71*.
Feldman, F. (1980), «Identity, Necessity and Events», en Block, 1980.
Feyerabend, P. K. (1963a), «Mental Events and the Brain»: *Journal of Philosophy, 60*; incluido en Borst, 1970, Rosenthal, 1971 y Lycan, 1990.
Feyerabend, P. K. (1963b), «Materialism and the Mind-Body Problem»: *Review of Metaphysics, 17*.
Fodor, J. (1974), «Special Sciences or the Disunity of Science as a Working Hypotesis»: *Synthese, 28*; incluido en Block, 1980.
Fodor, J. (1985), «Fodors Guide to Mental Representarions»: *Mind, 94*; incluido en Greenwood, 1991.
Fodor, J. (1987), «The Persistence of Attitudes», cap. 2 de *Psychosemantics*, MIT Press, Cambridge, Mass; incluido en Rabossi, s/f.
Greenword, J. D. (comp.) (1991), *The Future of Folk Psychology. Intentionality and Cognitive Science*, Cambridge University Press, Cambridge.
Hook, S. (comp.) (1960), *Dimensions of Mind*, New York University Press, New York.
Kim, J. (1972), «Phenomenal Properties, Psychophysical Laws, and the Identity Theory»: *Monist, 56*; incluido, en parte, en Block, 1980, bajo el título «Physicalism and the Multiple Realizability of Mental States».
Kim, J. (1982), «Psychological Supervenience as a Mind-Body Theory»: *Cognition and Brain Theory, 5*.
Kim, J. (1989), «The Myth of Non-Reductive Materialism»: *Proceedings and Addresses of the American Philosophical Association, 63*.
Kim, J. (1992), «Multiple Realization and the Metaphysics of Reduction»: *Philosophy and Phenomenological Research, 52*.

Kripke, S. (1971), «Identity and Necessity», en M. Munitz (comp.), *Identity and Individuation*, New York University Press, New York; incluido, en parte, en Block, 1980.
Kirpke, S. (1972), «Naming and Necessity», en D. Davidson y G. Harman (comps.), *Semantics of Natural Language*, Reidel, Dordrecht.
Lewis, D. (1966), «An Argument for the Identity Theory»: *Journal of Philosophy*, *63*; incluido en Rosenthal, 1971; Lewis, 1983.
Lewis, D. (1969), «Review of Putnam» (1967): *Journal of Philosophy*, *66*; incluido en Block, 1980.
Lewis, D. (1970), «How to Define Theoretical Terms»: *Journal of Philosophy*, *67*; incluido en Lewis, 1983.
Lewis, D. (1972), «Psychophysical and Theoretical Identifications»: *Australasian Journal of Philosophy*, *50*; incluido en Block.
Lewis, D. (1974), «Radical Interpretation»: *Synthese*, *23*; incluido en Lewis, 1983.
Lewis, D. (1980), «Mad Pain and Martian Pain», en Block, 1980; incluido en Lewis, 1983.
Lewis, D. (1983), *Philosophical Papers* I, OUP, Oxford.
Lycan, W. (comp.) (1990), *Mind and Cognition*, Blackwell, Oxford.
MacDonald, C. (1989), *Mind Body Identity Theories*, Routledge, London.
McGinn, C. (1982), *The Character of Mind*, OUP, Oxford.
Nagel, E. (1961), *The Structure of Science*, Harcourt, Brace and World, New York. V.e.: N. Miguez, Paidós, Barcelona.
Nagel, T. (1970), «Armstrong on the Mind»: *Philosophical Review*, *79*; incluido en Block, 1980.
Oppenheim, P. y Putnam, P. (1958), «Unity of Science as a Working Hypothesis», en H. Feigl, M. Scriven y G. Maxwell (comps.), *Minnesota Studies in the Philosophy of Science* II, University of Minnesota Press, Minneapolis.
Place, U. T. (1956), «Is Consciousness a Brain Process?»: *British Journal of Psychology*, *47*; incluido en Borst, 1970 y Lycan, 1990.
Place, U. T. (1960), «Materialism as a scientific hypothesis»: *Philosophical Review*, *69*; incluido en Borst, 1970.
Presley, C. F. (comp.) (1967), *The Identity Theory of Mind*, University of Queensland Press, Santa Lucia.
Putnam, H. (1960), «Minds and Machines», en Hook, 1960.
Putnam, H. (1967), «The Nature of Mental States», en W. H. Capitan y D. D. Merrill (comp.), *Art, Mind and Religion*, University of Pittsburg Press, Pittsburg. Publicado originariamente bajo el título «Psychological Predicates»; incluido en Borst, 1970; Rosenthal, 1971 y Lycan, 1990.
Rabossi, E. (comp.) (s/f), *Filosofía de la mente y ciencia cognitiva*, Paidós, Barcelona (en prensa).
Rorty, R. (1965), «Mind-Body Identity, Privacy and Categories»: *Review of Metaphysics*, *19*; incluido en Borst, 1970 y Rosenthal, 1971.
Rosenthal, D. M. (comp.) (1971), *Materialism and the Mind-Body Problem*, Prentice-Hall, Englewood Cliffs.
Shoemaker, S. (1981), «Some Varieties of Functionalism»: *Philosophical Topics*, *12*.

Smart, J. J. C. (1959), «Sensations and Brain Processes»: *Philosophical Review*, *68*; incluido en Borst, 1970 y Rosenthal, 1971.
Smart, J. J. C. (1961), «Further Remarks on Sensations and Brain Processes»: *Philosophical Review*, *70*; incluido en Borst, 1970.
Smart, J. J. C. (1962), «Brain Processes and Incorregibility»: *Australasian Journal of Philosophy*, *40*; incluido en Borst, 1970.
Smart, J. J. C. (1963), *Philosophy and Scientific Realism*, Routledge, London.
Smart, J. J. C. (1963), «Materialism»: *Journal of Philosophy*, *60*; incluido en Borst, 1970.
Smart, J. J. C. (1967), «Comments on the Papers», en Presley, 1967.
Stevenson, J. T. (1960), « Sensations and Brain Processes: A Reply to J. J. C. Smart»: *Philosophical Review*, *70*; incluido en Borst, 1970.
Stich, S. (1983), *From Folk Psychology to Cognitive Science*, MIT Press, Cambridge, Mass.
Valdés, M. (1980), «La objeción de Kripke a la teoría de la identidad mente-cuerpo»: *Dianoia*, *26*.

EL FUNCIONALISMO *

Manuel García-Carpintero

I. INTRODUCCIÓN Y DELIMITACIÓN DEL TEMA

Con «estado» designaremos esas entidades que aparecen como «factores causales» en nuestras explicaciones: entidades a las que suponemos el poder para causar y ser causadas. Son estados en este sentido tanto *acaecimientos* —cambios de duración relativamente corta: la caída de un rayo, la proyección de una película o la rotura de un freno— como *procesos* —cambios «prolongados»: la subida de la marea, la deriva continental o la desaparición de los dinosaurios— como *estados* propiamente dichos, en que el cambio no está presente —el estado de engrase del motor o la presencia de arena en la curva—. Entre los estados incluimos, en nuestra ontología cotidiana, los *estados mentales*. Los estados mentales incluyen también acaecimientos —la «visión» como en un *flash* de la solución al problema—, procesos —la consideración de todos los pros y los contras relevantes previa a la toma de una decisión importante— y estados propiamente dichos —la opinión de que las películas de John Huston son amenas—. Pues también a los estados mentales les suponemos el poder para causar y ser causados: aparentemente en el mismo sentido en que decimos que la caída de un rayo causó la destrucción del árbol, que la subida de la marea causó que la toalla se empapase y que el estado de engrase del motor causó la avería, decimos también que su sonrisa de satisfaccción la causa la comprensión de la solución del

* Agradezco a Fernando Broncano, Ramón Cirera, José Antonio Díez, Manuel Pérez y David Pineda su cuidadosa lectura de una versión anterior del artículo, así como sus comentarios y críticas, que han contribuido a mejorarlo. La investigación necesaria para confeccionarlo ha sido parcialmente financiada por la DGICYT a través del proyecto PB-90-0701-C03-03.

problema, que la consideración de los pros y los contras de la cuestión hizo que se matriculase finalmente en Medicina y que su opinión de que las películas de John Huston son amenas le llevó a ir esa tarde al cine.

Las expresiones con las que describimos estados mentales, como «Julia opina que las películas de John Huston son amenas» o «Arquímedes vio cómo Siracusa era destruida», presentan la siguiente estructura: una expresión con la que designamos al *sujeto* del estado, «Julia» y «Arquímedes»; una parte con la que designamos el *tipo* del estado, «opina» y «vio», y por último una parte con la que describimos el *contenido* del estado, «que las películas de John Huston son amenas» o «cómo Siracusa era destruida». El uso de la expresión «contenido» se apoya en la siguiente analogía. En el discurso directo —«Julia dijo: las películas de John Huston son amenas»— debemos ser fieles a las palabras que atribuimos al sujeto. Lo que se dice en este enunciado no sería literalmente verdad si Julia no hubiese utilizado exactamente las palabras que le atribuimos, si, por ejemplo, Julia hubiese hablado en catalán. En el discurso indirecto —«Julia dijo que las películas de John Huston son amenas»—, sin embargo, las condiciones son menos estrictas. Todo lo que debemos hacer para hablar con verdad es describir correctamente el *contenido* de las palabras de Julia, cualesquiera que éstas fuesen: *lo que* Julia dijo. Pues bien, la expresión «que las películas de John Huston son amenas» cumple el mismo papel en «Julia opina que las películas de John Huston son amenas» que en «Julia dijo que las películas de John Huston son amenas»: en ambos casos está ahí para describir el *contenido*, bien el de la opinión de Julia, bien el de las palabras que de hecho utilizó.

Así pues, el modo en que hablamos de la mente indica que atribuimos a los estados mentales contenido, al igual que se lo atribuimos a las expresiones lingüísticas. El contenido del enunciado catalán «Schlick i Wittgenstein tingueren moltes entrevistes a les quals acudí quasi sempre Waismann» es que *Schlick y Wittgenstein tuvieron muchas entrevistas a las que casi siempre asistió Waismann*. En virtud del contenido que tiene, una expresión lingüística —un enunciado— representa la realidad de un cierto modo. Dado su contenido, es decir, dado el modo en que el enunciado representa la realidad, y dado cómo de hecho es el mundo, un enunciado es verdadero o falso. Si los estados mentales tienen contenido es que también ellos *representan* la realidad de un cierto modo: el contenido de los estados mentales consiste en la especificación de los detalles de la representación que en ellos se hace de la realidad. Una opinión de que los dinosaurios se extinguieron en el Cretácico representa el mundo como conteniendo la extinción de los dinosaurios en un cierto período de tiempo; dado ese contenido, la opinión es ulteriormente verdadera o falsa en virtud de cómo, de hecho, sea el mundo. Un deseo de poseer un

yate representa la realidad como incluyendo al sujeto del deseo siendo poseedor de un yate en algún momento de tiempo futuro respecto de aquel en que se da el deseo; dado ese contenido, el deseo será ulteriormente «verdadero» o «falso» en virtud de cómo, de hecho, sea la realidad.

En el caso de estados mentales como los deseos decimos mejor «satisfecho» o «insatisfecho» que «verdadero» o «falso». Pero la relación entre las opiniones de las que decimos que son verdaderas y el mundo, y la relación entre los deseos de los que decimos que son satisfechos y el mundo son sustancialmente similares: ambas son relaciones de buen «ajuste». La diferencia, en afortunada caracterización de Searle (1983), consiste en la «dirección» del ajuste: en el caso de las opiniones verdaderas, el contenido se acomoda al mundo; en el de los deseos satisfechos, el mundo se acomoda al contenido del deseo. De modo menos misterioso, podríamos decir con pleno espíritu funcionalista que las opiniones con contenidos verdaderos han sido producidas por los aspectos del mundo representados en el contenido, mientras que los deseos con contenidos satisfechos producen ellos los aspectos del mundo representados en el contenido.

Este carácter representacional que suponemos a los estados mentales y que recogemos al caracterizar su contenido no lo comparten muchos otros estados; la caída de un rayo, la deriva continental o la presencia de arena en la curva no parecen representar nada, salvo, quizás, metafóricamente hablando. Al menos un filósofo contemporáneo, Franz Brentano, consideró a tal carácter representacional de los estados mentales su rasgo distintivo, la esencia que los separa de otros tipos de estados. Brentano se refería a ese aspecto esencial de los estados mentales como su *intencionalidad*: los estados mentales, necesariamente, *tienden hacia* otros estados, representan la realidad (correcta o incorrectamente, eso se deja al albur de la realidad misma) como conteniendo ciertos otros estados —aquellos que constituyen su contenido—. La clave para la comprensión del concepto de lo mental reside, si Brentano tiene razón, en elucidar esta noción de *contenido* que hasta aquí hemos presentado intuitivamente, el carácter *representacional* o *intencional* de los estados mentales. Una teoría filosófica de la mente, por consiguiente, tiene como tarea primera la de dar cuenta de la intencionalidad de lo mental.

II. LA CONCEPCIÓN CARTESIANA DE LA MENTE

Quizás sea la cartesiana la concepción filosófica de lo mental que, a primera vista, mejor acuerda con nuestras intuiciones cotidianas; sin duda lo hace con nuestros anhelos. Descartes se refiere a lo que nosotros llamamos *estado mental* con la expresión «pensamiento», y dice: «Con la

expresión “pensamiento” entiendo todo aquello que sucede dentro de nosotros de lo que somos conscientes, en la medida en que tenemos consciencia de ello» (Descartes, *Principios de Filosofía*, i, §9.) La mente es fundamentalmente consciencia, y la consciencia un cierto tipo de conocimiento privilegiado «de lo que sucede dentro de nosotros». Si yo tengo un estado mental, digamos la opinión basada en el testimonio de mis sentidos de que hay una esfera roja del tamaño aproximado de una manzana en cierta posición ante mí, entonces, de acuerdo con la concepción cartesiana, yo soy consciente de tener tal estado; y este «ser consciente» implica, como mínimo, que *me sé* tenedor de una opinión con tal contenido. Este conocimiento no es del tipo del que nos consideramos poseedores, digamos, sobre la no existencia de vida en Marte; el conocimiento consciente es un conocimiento *cierto*, con una certidumbre que no deja resquicio alguno a la duda.

La concepción cartesiana de la mente resulta ciertamente plausible a primera vista. Algo así debe subyacer a la creencia generalizada en la existencia de un abismo cualitativo entre lo mental y lo físico, que está detrás de muchas opiniones religiosas: la mente es esencialmente consciencia, y los estados conscientes son necesariamente distintos de estados no conscientes como los estados físicos de nuestro organismo, más específicamente los de nuestro cerebro. El propio Descartes, como es bien sabido, construyó a partir de su concepción de lo mental un argumento cuidadoso para demostrar la existencia de una «distinción real» entre la mente y el cuerpo, distinción que pretende justificar la posibilidad de la existencia separada de la primera.

Un examen más detenido, sin embargo, basta para revelar alguna perplejidad. Un estado consciente es un estado de conocimiento, y de conocimiento cierto, pero *¿de qué* ofrece conocimiento un tal estado? Obsérvese que Descartes se refiere a los pensamientos como «lo que sucede *dentro* de nosotros». Esto puede parecer inocuo, debido a que con «pensamiento» nos referimos tanto al acto de pensar, que sí sucede «dentro» de nosotros, como a lo pensado, al contenido del pensamiento. Pero un poco de reflexión revela que no es inocuo, pues Descartes se refiere también a lo pensado. Lo que yo conozco cuando, por ejemplo, me considero percibiendo una esfera roja acercándose a mí es la esfera roja en movimiento, y «eso» parece estar «fuera». Lo que lleva a Descartes a describir también *lo pensado* como ocurriendo «dentro de nosotros» es que mi percepción podría ser sólo aparente; yo podría estar padeciendo una alucinación o una ilusión perceptiva, y en ese caso el rojo de la esfera, su esfereidad, su posición espacial y su movimiento en el tiempo podrían ser todos ellos «meras apariencias», «fenómenos». Por otra parte, «internamente» yo no puedo distinguir si mi percepción es verídica o meramente aparente: tanto las percepciones verídicas como las aparentes «se experimentan» fenoménicamente igual. Es fácil concluir de estas consi-

deraciones que aquello que yo «inmediatamente» conozco cuando soy consciente, tanto cuando realmente percibo como cuando padezco alucinaciones, es un conjunto ordenado de meras características fenoménicas, meras apariencias. En cualquier caso, este es el punto de vista de Descartes. Los objetos conocidos en los estados de consciencia son fenómenos; son éstos los pensamientos que están «dentro» de nosotros, aunque «aparezcan» fuera. Los modos o elementos de las apariencias (el rojo, la esfereidad, el espacio en que la esfera roja se ubica, el tiempo a lo largo del cual se mueve) son las «ideas» de Descartes y Locke, los *qualia* de los filósofos contemporáneos.

Si, como el cartesiano piensa, los estados conscientes son estados en que tenemos certidumbre de ciertas presencias concurrentes con los estados; si somos infalibles respecto de las características de lo que conocemos conscientemente y nadie está nunca en situación de corregir nuestra opinión sobre la naturaleza de tales presencias concurrentes a los estados de consciencia, las presencias en cuestión tienen ciertamente que ser fenómenos. Yo no puedo saber con certeza que, concurriendo con lo que tomo por una percepción de una esfera roja moviéndose hacia mí, está presente en el espacio «real» en torno a mí una «real» esfera «realmente» roja «realmente» moviéndose en el tiempo «real». Descartes es famoso por haber inventado la posibilidad de una mayúscula alucinación, recurriendo a la figura del Genio Maligno. Incluso si esta extravagante invención no fuese una mera posibilidad, sino la verdad misma, si mis estados mentales conscientes (esto desde luego es una redundancia para el cartesiano) fuesen después de todo el producto de las maquinaciones de tal Genio Maligno, mi consciencia sería igualmente fidedigna, pues las apariencias que a través de ella conozco serían justamente las que son.

La perplejidad que esto produce (por más familiares que nos resulten las reflexiones precedentes) proviene de la radical separación que se establece entre la naturaleza del mundo objetivo y la naturaleza de los «universos interiores» de que propiamente hablando (según esta concepción) somos conscientes. El modo en que hablamos no refleja esa disociación, sino todo lo contrario. Sería natural pensar que las palabras tienen el mismo significado cuando se utilizan para hablar directamente de la realidad que cuando se utilizan en estilo indirecto, para indicar el contenido de un estado mental. La concepción cartesiana de la mente, empero, desmiente esta intuición, o tergiversa su prístino sentido. El enunciado «hay una esfera roja del tamaño aproximado de una manzana en cierta posición ante mí» «trata» de lo que ocurre, objetivamente, ante mí: pues si yo estoy padeciendo una alucinación, y no hay ninguna esfera ante mí, ese enunciado es falso. Pero las palabras «hay una esfera roja del tamaño aproximado de una manzana en cierta posición ante mí», en «opino que hay una esfera roja del tamaño aproximado de una

manzana en cierta posición ante mí», no «tratan» del mundo externo. «Rojo» y «esfera», en «hay una esfera roja ante mí» y en «opino que hay una esfera roja ante mí», no tienen el mismo significado. Una vez acepto la concepción cartesiana de la mente, he de convenir en que la rojez que caracteriza el contenido de mi estado es distinta de cualquier rojez que pudiera adornar objetivamente a las esferas reales, pues la primera podría existir aunque no existiera la segunda: que no hubiera nada realmente rojo en el mundo no afectaría según el cartesiano un ápice a la existencia de opiniones cuyo contenido representa esferas rojas.

Una vez que reparamos en este aspecto de la concepción cartesiana de la mente comienzan a asaltarnos todo tipo de dudas, que ponen en cuestión si realmente este análisis constituye una buena teoría de lo mental. Aparece en primer lugar la cuestión de las «otras mentes»: la certidumbre que caracteriza lo mental en la perspectiva cartesiana sólo existe respecto de los contenidos de *mis propios* estados mentales; cuál sea el de los demás, si es que siquiera cabe atribuírselos a los demás en esta concepción, es cosa tan dudosa, o más, de lo que pueda serlo la naturaleza del mundo externo. La cuestión de cuáles sean las presencias fenoménicas concurrentes con los estados conscientes de los otros cuando describen algo como rojo, si alguna hay (quizás los otros son robots diseñados por un científico muy sabio, y en realidad no son conscientes de nada) ha de permanecer irresoluble. En segundo lugar, no hace falta aceptar los principios de la psicología freudiana para encontrar múltiples ejemplos de estados que guardan suficientes analogías con los más característicamente mentales como para contarlos entre éstos y, sin embargo, no son estados conscientes. Los rasgos de carácter, como el valor o la indiscreción, y la mayoría de los procesos que garantizan que cada día conduzca sano y salvo mi vehículo hasta la universidad son buena muestra de ello.

Y resta finalmente la más grave dificultad. Los estados mentales son estados, como dijimos al comienzo, porque les suponemos el poder para causar y ser causados. Entre los acaecimientos presuntamente causados por ellos hay acaecimientos físicos: mi opinión de que hay una esfera roja ante mí, junto con mi deseo de coger una esfera roja, son la causa de que mueva mi brazo como lo muevo en dirección a la esfera. Por otra parte, nosotros suponemos que el dominio de lo físico es un dominio causalmente cerrado: si retrotraemos la historia causal de un acaecimiento físico, sólo encontraremos otros acaecimientos físicos en nuestro camino. Así, el movimiento de mi brazo, un acaecimiento físico, está causado por ciertos impulsos nerviosos afectando a ciertos músculos, impulsos nerviosos que están causados por ciertos impulsos nerviosos en el córtex, que están causados a su vez, entre otras cosas, por ciertos impulsos nerviosos con origen en la retina, que a su vez están causados por la luz reflejada por la esfera, etc. Ahora bien, las tesis de la eficacia causal de la

mente y de la completitud causal de lo físico son difícilmente conciliables, supuesta la concepción cartesiana de la mente: ¿cómo puede ser esa presencia fenoménica concurrente con mi percepción de la esfera roja de la que tengo un conocimiento infalible e incorregible un estado físico, por ejemplo un estado de mi cerebro? El propio Descartes sugirió abandonar la segunda tesis. En algún momento, en la historia anterior, habría que introducir la afección de algo físico por algo no físico, mental. Sus propuestas a este respecto no resultaron muy convincentes —ya desde el momento en que se expusieron—. Es una opinión generalizada (aunque no unánime) que la actitud más coherente para el cartesiano es abandonar la primera tesis (la eficacia causal de los estados conscientes), aceptando que la mente es *epifenoménica*: los estados mentales son epifenómenos carentes del poder para causar. Además de dudosamente inteligible, sin embargo, esta tesis contrasta con los hechos: existe una correlación sistemática entre los cambios en el cerebro y los cambios en la mente. ¿No habremos ido demasiado deprisa al dar por buena la concepción cartesiana de la mente? O ¿no habremos extraído consecuencias erróneas del acceso cognoscitivamente privilegiado a la naturaleza de nuestros propios estados mentales conscientes?

III. LA ALTERNATIVA CONDUCTISTA

El *conductismo* responde positivamente a estas preguntas. En rigor, conviene distinguir el *conductismo metodológico* del *conductismo lógico*. El primero (la doctrina popular entre los psicólogos en los años cincuenta y sesenta, gracias a Watson y Skinner) no niega la existencia de estados mentales sólo accesibles a la intuición del sujeto que está en ellos con las características que el cartesiano les atribuye, ni siquiera su carácter de paradigma de lo mental. Se limita a recomendar a la psicología la prohibición de teorizar sobre ellos; pues una psicología que no se atuviere a tal prohibición, habiendo de descansar entonces como fuente exclusiva de evidencia empírica en los dictámenes obtenidos por sus sujetos (típicamente, el propio psicólogo) mediante la introspección, única autoridad cada uno de ellos sobre la corrección o incorrección de tales dictámenes, sería una disciplina apartada del paradigma de control intersubjetivo que hacen a otras merecedoras del calificativo de *científicas*. El conductismo lógico, por otro lado, es una doctrina mucho más radical. Con la característica osadía del filósofo, el conductista lógico sostiene que la teoría cartesiana de lo mental es fruto de una confusión; que el concepto cartesiano de lo mental es incoherente, por lo que no sólo no hay, sino que *no puede haber*, estados mentales con las características que el cartesiano les atribuye. Rudolf Carnap (1932-33) Gilbert Ryle (1949) y el Wittgenstein del segundo período (1958) son ejemplos de este tipo de filóso-

fo[1]. Cada uno hace un diagnóstico distinto de las contradicciones en que incurre la concepción cartesiana de la mente; y la teoría que cada uno tiene que ofrecer sobre lo mental difiere también, en detalles a veces sustanciales. Pasaremos aquí sin embargo por alto tales detalles para encontrar un núcleo reconociblemente común.

Para el conductista (suprimiré el calificativo, pues en adelante sólo hablaré del conductismo lógico), cuando mencionamos estados mentales, no nos podemos estar refiriendo a estados de los que su sujeto, y sólo él, tiene un conocimiento privilegiado, lo que conlleva que su contenido concierne a entidades peculiarmente no «objetivas», sino «internas». Lo que hacemos es simplemente describir su conducta. Naturalmente, no la conducta que de hecho lleva a cabo; pues puede ser muy cierto que alguien es indiscreto, o que tiene la opinión de que hay una esfera roja ante él, sin que de hecho esté haciendo nada. Lo que describimos más bien es la conducta que *podría* llevar a cabo. Conceptualmente, las atribuciones de estados mentales son similares a la atribución de *disposiciones*. Cuando decimos que algo es *soluble* no estamos diciendo que de hecho se esté disolviendo, ni siquiera que se haya nunca de disolver. Estamos haciendo una afirmación condicional que se expresa propiamente en subjuntivo: decimos que se *disolvería* si se dieran ciertas condiciones (por ejemplo, si se *pusiera* en agua).

Hay diferencias entre disposiciones como la solubilidad y los estados mentales, ciertamente. Una bastante notoria consiste, en afortunada expresión de Ryle, en que mientras la solubilidad es una disposición de una *única vía* (es decir, tiene un único tipo de manifestación), los estados mentales son disposiciones de *múltiples vías*. La opinión de que hay una esfera roja ante mí consiste en que yo *proferiría* las palabras «hay una esfera roja ante mí» si se me *preguntase* qué hay ante mí (y se dieran otras condiciones que conviene notar ya: que yo *hablase* castellano; que *entendiese* la pregunta: que me *apeteciese* contestarla, etc.); o en que *movería* el brazo de un cierto modo, si se me *pidiese* que tocase el objeto rojo más cercano (y, de nuevo, se diesen otras condiciones, como que yo *no pensase* que la esfera está eléctricamente cargada, etc.); y existen múltiples otras manifestaciones conductuales posibles, constitutivas todas ellas según el conductista de la opinión mencionada.

Hay otras diferencias más sutiles entre disposiciones como la solubilidad y los estados mentales. Wittgenstein insistió en que, a diferencia de otras disposiciones, los estados mentales tienen naturaleza *normativa*.

1. He creído apropiado, dados los objetivos de este trabajo, prescindir de algunos usos académicos. No justifico, así, salvo aquellos puntos de vista directamente relacionados con el desarrollo del tema. El lector está advertido, naturalmente, de que la mayoría de las afirmaciones que se hacen son controvertibles, y algunas fuertemente controvertidas. La tesis de que el Wittgenstein del segundo periodo es un conductista lógico cuenta entre ellas. Menos controvertida es la similar atribución a Carnap, pero véase Cirera (1990).

Decir de un objeto que es soluble implica que en determinadas circunstancias se darán ciertos estados; pero si en las circunstancias en cuestión los estados no se dieran, no diríamos que ha ocurrido algo «incorrecto», sino más bien que ha ocurrido algo inesperado, o milagroso, o inexplicable. Decir que alguien tiene la opinión de que hay un abismo ante él sí implica, en cambio, que si el sujeto en cuestión, no deseando suicidarse, y no habiendo modificado su opinión sobre la presencia del abismo, acto seguido echa a correr hacia adelante, lo que está haciendo es no sólo inesperado, inexplicable o milagroso, sino «incorrecto» o «inapropiado» o «irracional» también. Los estados mentales son según Wittgenstein disposiciones de un tipo peculiar particularmente en cuanto a su normatividad, que consiste en que tenerlos implica la existencia de una distinción entre manifestaciones *correctas* y manifestaciones *incorrectas*, a diferencia de lo que ocurre con otras disposiciones (para las que tenerlas tan sólo implica la distinción entre manifestaciones probables y manifestaciones improbables). No discutiremos aquí la explicación *social o* comunitaria que Wittgenstein creía ser preciso dar de esta normatividad. (Ni la relación de la cuestión de la normatividad con su propia razón para creer que el concepto cartesiano de estado mental es incoherente, esto es, el célebre argumento contra la posibilidad de un «lenguaje privado».)

El conductismo está aparentemente libre de las dificultades de la concepción cartesiana de la mente que hemos mencionado antes (como no podía ser menos, por cuanto es una respuesta crítica a esa concepción). Desde un punto de vista conductista, no hay ningún abismo entre mente y mundo; dicho al modo analítico, en términos lingüísticos, las palabras «tratan de lo mismo» cuando se utilizan para describir la realidad («hay una esfera roja ante mí») y cuando se utilizan para describir la mente («opino que hay una esfera roja ante mí»). En el primer caso, enunciamos la existencia de una cierta situación externa; en el segundo, enunciamos la existencia de una disposición, cuyas manifestaciones son relativas *a la misma situación externa* aseverada por el enunciado anterior. En consecuencia, no existe el problema de las «otras mentes»: no hay ninguna dificultad en justificar la atribución de estados mentales a otros. O, mejor dicho, existe tanta dificultad en justificar estas afirmaciones como pueda existir en justificar las afirmaciones que hacemos mediante condicionales en subjuntivo (de las que la ciencia está llena): «se *disolvería* si *fuera puesto* en agua». Y tampoco hay dificultad alguna en incluir entre los estados mentales estados de los que no tenemos consciencia, o en atribuir estados mentales a entidades de las que dudamos que tengan consciencia, como algunos animales. La cuestión de la eficacia causal de la mente tampoco presenta, aparentemente, dificultades para el conductismo. La eficacia causal de los estados mentales, entendidos como disposiciones, viola la completitud causal del mundo físico en la misma medida en que lo hace suponer a la solubilidad de un terrón de

azúcar eficacia causal. (Como se indicará después, algunos filósofos tienen serias dudas sobre la eficacia causal de las disposiciones. Estas dudas, sin embargo, son completamente generales, y por consiguiente deben ser distinguidas de las dudas específicas sobre el papel causal de las propiedades mentales suscitadas por el cartesianismo.)

Naturalmente, todo esto tiene un precio; después de todo, la concepción cartesiana no es sólo el producto de una mente inventiva. El precio es en algún caso (particularmente, el de la dificultad de explicar la naturaleza de los estados mentales conscientes, pero también ciertas dudas sobre la eficacia causal de las disposiciones) compartido por el conductismo con su heredero natural, el funcionalismo, y evitaremos repeticiones discutiendo la cuestión en relación a esta doctrina. En otros se trata de dificultades específicas al conductismo, entre ellas las dos que hicieron que la mayoría de los filósofos lo abandonase durante los años sesenta. Se trata de las siguientes.

Cuando ejemplificamos antes el hecho de que esas disposiciones que, según el conductista, son los estados mentales, a diferencia de disposiciones tales como la solubilidad o la elasticidad, están constituidas por múltiples manifestaciones, quizás el lector advirtió que en la especificación de esas múltiples manifestaciones intervinieron *otros estados mentales*. Una opinión de que hay una esfera roja ante mí consiste en que yo *proferiría* las palabras «hay una esfera roja ante mí» si se me *preguntase* qué hay ante mí, yo *hablase* castellano, *entendiese* la pregunta, me *apeteciese* contestarla; en que *movería* el brazo de un cierto modo si se me *pidiese* que tocase el objeto rojo más cercano, no *pensase* que la esfera está eléctricamente cargada, etc. Los verbos en cursiva hacen referencia, de modo implícito o explícito, a otros estados mentales. Estos ejemplos no son accidentales. Si el lector trata de especificar las manifestaciones conductuales en que podría razonablemente consistir un estado mental, concediendo momentáneamente al conductista la corrección de su análisis, descubrirá que no hay ninguna posibilidad de enunciar las condiciones en que se darían los comportamientos pertinentes al caso sin hacer mención en ellas de otros estados mentales. No hay ninguna posibilidad, naturalmente, en la medida en que queramos que el análisis pueda al menos aspirar a parecer correcto. Sería manifiestamente incorrecto contar la manifestación «el individuo en cuestión *movería* el brazo de un cierto modo, si se le *pidiese* que tocase el objeto rojo más cercano» como una condición necesaria en un análisis de *opinión de que hay una esfera roja ante uno*, pues, claramente, alguien podría muy bien estar en ese estado mental y no tener la disposición descrita: alguien, por ejemplo, que *creyese* que tocar una esfera roja produce instantáneamente la muerte[2].

2. Esta objeción fue elaborada por varios filósofos. Una versión influyente puede verse en Geach, 1992, §4. Ni que decir tiene que conductistas como Ryle eran conocedores de la dificultad. Les

Para definir un estado mental en términos de disposiciones conductuales, por consiguiente, es preciso incluir en la descripción de las condiciones en que se darían los comportamientos pertinentes la referencia a otros estados mentales. El problema que esto plantea al conductismo es claro. El conductista cree que los conceptos mentales son conceptos de disposiciones a conducirse de ciertos modos en ciertas circunstancias. Si esto es así, los estados mentales deben ser definibles sin remanente alguno en términos de tales disposiciones. Pero esta condición no puede ser satisfecha si las condiciones incluyen referencias a otros estados mentales.

Nos referiremos a este primer problema con el que tropieza el conductismo como el del *holismo de lo mental*: el concepto de un estado mental parece estar inextricablemente vinculado al de otros estados mentales, de modo que los conceptos mentales forman un todo interrelacionado.

La segunda objeción al conductismo se apoya en la inevitable intuición de que los estados mentales no están *constituidos* por disposiciones a la conducta en ciertas circunstancias externas, sino que tales disposiciones son sólo un *resultado* de los mismos. Esta objeción se puede formular elaborando contraejemplos intuitivos basados en ella, y el más famoso es uno debido a Hilary Putnam[3]. Putnam nos pide que imaginemos una comunidad de superespartanos. Los superespartanos son individuos que han conseguido suprimir toda manifestación del dolor distinta de su manifestación verbal. Por consiguiente, no es verdad de estos individuos que si se les extrajese una muela sin anestesia gritarían, gemirían, llorarían, etc. (Disposiciones estas con las que el conductista intentaría definir el dolor.) Las únicas disposiciones relacionadas con el dolor que estos individuos comparten con los seres humanos normales son disposiciones a la conducta verbal; por ejemplo, la disposición a decir que sienten dolor cuando se les extrae una muela sin anestesia (o a utilizar las palabras «tengo dolor» cuando se les extrae una muela, y se da además la circunstancia de que saben castellano). Por lo demás, es posible comprobar que la extracción de una muela causa en estos individuos un proceso nervioso, desde el lugar de la extracción hasta el córtex cerebral, idéntico al que causa en un ser humano normal. Intuitivamente, estos individuos *tienen dolor*; la única diferencia con los individuos normales radica en que los superespartanos, cuando tienen dolor, tienen también un estado que los individuos normales no tienen, o no tienen en ese grado, a saber, el deseo en grado superlativo de suprimir cualquier manifestación no verbal del dolor.

diferenciaba de sus oponentes el peso que le concedían en el balance global de los pros y los contras de su teoría, y el optimismo respecto de la posibilidad de solucionarla.

3. Véase Putnam (1975a).

Hasta aquí no tenemos aún una objeción al conductismo, sino a lo sumo una elaboración del problema anterior (los estados mentales sólo pueden verse como disposiciones a la conducta si las condiciones en que se manifestarían hacen referencia a otros estados mentales): el conductista podría dar cuenta de nuestro juicio intuitivo apoyándose en las disposiciones a la conducta verbal aún compartidas por el superespartano y el individuo normal. Ahora bien, Putnam nos pide que imaginemos después una comunidad de super-superespartanos. Estos son idénticos en todo a los superespartanos, salvo en que además han suprimido las manifestaciones verbales del dolor. No tienen, por tanto, ninguna de las manifestaciones conductuales con las que el conductista identifica el dolor. Sin embargo, parece que nuestro juicio intuitivo es el de que tales individuos son al menos concebibles, y que tendrían dolor. Como se dijo al comienzo, la fuente de este juicio, que el ejemplo meramente revela, es la firme creencia intuitiva de que los estados mentales son sólo *causa* de las disposiciones con que el conductista pretende identificarlos. Por ser sólo causa de ellas, es posible concebir la presencia de los estados mentales aun en ausencia de sus resultados habituales, esas disposiciones a la conducta en que de acuerdo con el conductista consistirían; y tal posibilidad manifiesta lo erróneo de identificar estados mentales y disposiciones conductuales.

IV. FUNCIONALISMO COMPUTACIONAL Y FUNCIONALISMO ANALÍTICO

El funcionalismo es una teoría sobre los conceptos de estados mentales elaborada por Hilary Putnam en una serie de artículos publicados originalmente durante los años sesenta [4]. La teoría puede verse como un desarrollo de temas enfatizados por el conductismo, resultado de considerar objeciones a esta doctrina como las que acabamos de formular. Putnam presentó su teoría mediante la analogía de los programas de ordenador (particularmente los favoritos de los lógicos, las «máquinas» de Turing), pero la versión más general debida a David Lewis suele ser preferida hoy.

Una *descripción funcional* es la descripción de un proceso causal: la descripción de cómo una serie de *inputs* convenientemente identificados dan lugar a una serie de *outputs* a través de un cierto tipo de proceso. La descripción de una cadena de montaje de un cierto modelo de vehículo se ajusta a esta caracterización: se trata de la descripción de cómo una serie de inputs, los elementos a partir de los cuales se construye el vehí-

4. A partir de «Minds and Machines». Como éste, la mayoría de ellos están recopilados en sus *Philosophical Papers*, vol. 2. El artículo en que se presenta la versión estándar de la teoría es Putnam (1975b).

culo, dan lugar a un vehículo finalizado a través de un proceso de ensamblaje que la descripción caracteriza con detalle. El maravilloso libro *Cómo funcionan las cosas* está repleto de descripciones funcionales[5]; por ejemplo, la descripción del sistema de cambio de marchas de un vehículo, comenzando con las diferentes posiciones de la palanca de cambios y del pedal del embrague (los *inputs*), y finalizando con la transmisión del movimiento del motor a las ruedas (el *output*). Un programa de ordenador, Word Perfect 5.1 por ejemplo, es, característicamente, una descripción funcional muy compleja. Será conveniente contar con un ejemplo simple de descripción funcional para hacer más vívidas las nociones y aseveraciones que se propondrán después. Se trata de la descripción del proceso que sigue una máquina expendedora de billetes de autobús; llamaremos F a la descripción que sigue.

La máquina admite monedas de cien y cincuenta pesetas, y está en uno de dos estados internos, S_1 y S_2. Cuando, estando en S_1, se introduce en ella una moneda de cien, da un billete de autobús y sigue en S_1; cuando estando en S_1 se introduce en ella una moneda de cincuenta, pasa a S_2; cuando estando en S_2 se introduce en ella una moneda de cien, da un billete de autobús, cincuenta pesetas, y vuelve a S_1; cuando estando en S_2 se introduce en ella una moneda de cincuenta, da un billete de autobús y vuelve a S_1.

Esta simple descripción funcional comparte con otras la característica de que en ella se hace referencia a ciertos estados intermedios en el proceso que lleva de los inputs (la introducción de las monedas) a los outputs (la entrega del billete y el cambio que corresponda), a los que aquí se hace referencia con «S_1» y «S_2». La descripción da poca información sobre ellos, pero esta información es suficiente para definirlos: S_1 es el estado de una máquina del tipo descrito tal que, cuando estando en él se recibe una moneda de cien, se da un billete y se continúa en él, y cuando estando en él se recibe una moneda de cincuenta, se pasa a S_2. S_2, por su parte, es el estado tal que, cuando se está en él y se recibe una moneda de cien, se devuelve una moneda de cincuenta, se da un billete de autobús y se pasa a S_1; cuando, por otro lado, estando en él se recibe una moneda de cincuenta, se da un billete de autobús y se pasa a S_1. Ambas características son muy generales y poco informativas, pero son distintivas, en el sentido de que acotan suficientemente los estados en cuestión, los distinguen de otros.

La descripción misma, F, es poco específica, en el sentido en que la condición «se entra a trabajar antes del mediodía» es menos específica que la condición «se entra a trabajar de nueve a diez»: más objetos, presentando más diferencias entre sí, cumplen F que otra descripción más es-

5. D. Macaulay, *The Way Things Work*, Houghton Mifflin Co., Boston, 1988 (v.e.: Anaya, Madrid).

pecífica que podríamos dar del proceso que sigue una máquina expendedora de billetes. En especial, dos objetos pueden cumplir F aunque los mecanismos que en cada uno de ellos constituyen el estado S_2 sean muy distintos entre sí. Los mecanismos en cuestión llevan a cabo, en ambos casos, las tareas características de S_2, del modo caracterizado por F; pero las llevan a cabo de modos muy distintos entre sí. Sin embargo, el que la descripción F sea muy poco precisa no significa que sea vacua. No lo es, porque no cualquier objeto la satisface. Una hucha corriente, por ejemplo, no la satisface. Una descripción funcional es la descripción de un proceso causal; poco detallada, pero aun así la descripción de un proceso causal. Para que un objeto satisfaga una descripción funcional tiene que ser posible identificar estados del objeto con los estados postulados en la descripción funcional, de modo tal que cualesquiera *inputs* posibles admitidos por la descripción *causarían* los *outputs* especificados en la descripción *a través de un proceso con las características del especificado* en la descripción. Una hucha corriente no satisface F, no sólo porque no daría billetes de autobús si se introdujesen en ella monedas de cien y de cincuenta en cualquier secuencia, sino también porque no es posible identificar estados de la hucha que lleven a cabo los papeles genéricos pero bien definidos asignados a S_1 y S_2 en F: no hay dos estados de la hucha tales que estar en el primero causaría que la hucha pasase a estar en el segundo si se introduce una moneda de cincuenta, tales que estar en el segundo causaría que la hucha diese un billete de autobús y pasase a estar en el primero si se introdujese en ella una moneda de cincuenta pesetas, etc. Cuando un objeto satisface una descripción funcional se dice que el objeto la *realiza*, la *implementa* o la *pone por obra*; y también de aquellos estados internos suyos que corresponden a los estados intermedios postulados en la descripción que los *realizan*, *implementan* o *ponen por obra*. Una hucha corriente carece de estados internos capaces de realizar los estados S_1 y S_2 descritos por F.

Un programa de ordenador, como se dijo anteriormente, es una descripción funcional, mucho más compleja, pero que comparte con F las características mencionadas: también incluye estados internos; también es cierto que los estados internos pueden ser completamente especificados a partir de la descripción; también es verdad que la descripción caracteriza de un modo menos informativo de lo que podría hacerse aquellos objetos que la satisfacen, particularmente los estados internos de esos objetos que corresponden a los estados intermedios postulados en el programa; y, por último, también es verdad que pese al carácter poco específico de la descripción y de la caracterización de los estados intermedios, no cualquier objeto la satisface. Para que un objeto lo haga, el objeto debe tener una cierta *estructura funcional*, aquella especificada por el programa: debe ser posible identificar estados suyos susceptibles de ordenarse en las secuencias causales indicadas en el programa. La silla en que

estoy sentado, por ejemplo, es incapaz de satisfacer la descripción funcional a que llamamos Word Perfect 5.1 [6].

Un programa de ordenador es la descripción funcional de un proceso que, en muchos casos, consiste en el desarrollo de tareas como aquellas que se consideran características de lo mental. La similitud ha estimulado el programa de investigación conocido como *inteligencia artificial*: desarrollar programas de ordenador (y objetos que los realicen) capaces de llevar a cabo tareas cada vez más similares a aquellas que suponemos requieren una mente. La similitud ha llevado aún más lejos a algunos: suponer que describir una mente, la mente humana por ejemplo, o la de los animales a los que les suponemos una mente, es formular una descripción funcional capaz de explicar aquellas actividades que parecen requerir una mente. El búho, por ejemplo, parece ser capaz de ubicar sonidos en el espacio tridimensional, pues es mediante ese recurso que parece poder cazar sus presas en la oscuridad. Desde el punto de vista sugerido por la analogía computacional, la tarea del biólogo que pretende explicar la percepción auditiva de los búhos sería formular una descripción funcional que indique cómo, a partir de los *inputs* relevantes (las propiedades de los sonidos que llegan a los órganos auditivos externos de los animales), se pueden producir los *outputs* que nos interesan (una conducta sensible a la procedencia de los sonidos en el espacio externo). Como ya sabemos, una descripción de este tipo no sería una mera estipulación, sino que, para que sea cierto que el búho la satisface, debe ser posible identificar estados neurológicos del animal con los estados postulados en la misma, de modo que los estados neurológicos sigan la secuencia causal especificada en la descripción [7].

6. Insisto en esto porque adversarios del funcionalismo como Noam Chomsky y John Searle han defendido que el que un objeto satisfaga o no una descripción funcional es algo que se estipula. Debe estar claro que esto puede ser verdadero de un sentido de «descripción funcional» específico a esos autores, pero es trivialmente falso del caracterizado aquí. Y también del usualmente supuesto. Si el que un objeto realice Word Perfect 5.1 pudiera estipularse, el lector podría conseguir un procesador de textos muy barato, «estipulando» que su bolígrafo lo hace. Searle (1992) admite que es necesario que un objeto tenga una cierta estructura causal, para que pueda decirse de él que realiza una descripción funcional. El bolígrafo, por tanto, no realiza Word Perfect. Pero insiste en que no es suficiente. Además, debe darse una «estipulación». Searle es libre de introducir los términos que crea conveniente, con el sentido que le parezca apropiado. Pero debe notarse que, cuando critica el funcionalismo, lo que critica es el funcionalismo que él ha caracterizado. En la versión que estamos presentando aquí (que es, por lo demás, la que el resto de los filósofos supone), es suficiente y necesario para que un objeto satisfaga una descripción funcional que el objeto tenga la estructura causal correspondiente a la descripción funcional. No es precisa ninguna ulterior estipulación.

7. Otra objeción al funcionalismo que se revela totalmente infundada cuando se atiende a esto es la de que una psicología funcionalista «no habría de prestar atención» a los datos biológicos y neurológicos, en contra de lo que la práctica de los psicólogos reales manifiesta. En efecto, si es una descripción funcional la que el psicólogo de los búhos busca, ciertamente la descripción misma será más abstracta que, por ejemplo, una descripción neurológica. (En el sentido de que la misma descripción podría ser implementada por un proceso muy distinto, no neurológico sino, por ejemplo, electrónico.) Pero como, para ser verdadera del búho, la descripción ha de ser realizable por el búho, y como en el

La propuesta de Putnam, el funcionalismo original, justificaría conceptualmente este programa de investigación en psicología que tan importantes frutos ha dado ya. La propuesta de Putnam es considerar que un estado mental es simplemente uno de los estados intermedios especificados por una descripción funcional capaz de dar cuenta de comportamientos detrás de los cuales característicamente suponemos una mente. Esta es una propuesta revisionista: no se trata de pretender que el concepto habitual de lo mental sea funcional, sino más bien de proponer que se entienda así. No cabe otro modo de entenderla cuando pensamos en las descripciones funcionales pertinentes bajo el modelo de los programas postulados por los psicólogos cognitivos para explicar la percepción del color o la comprensión del lenguaje. Llamemos *funcionalismo computacional* a esta versión de la doctrina. La versión más general del funcionalismo debida a Lewis permite ir más allá, y defender el funcionalismo no como una propuesta revisionista, sino como un análisis de nuestro concepto de lo mental, capaz de rivalizar con el análisis conductista o el cartesiano; denominaremos *funcionalismo analítico* a esta segunda versión [8].

Lewis propone concebir las teorías científicas como descripciones funcionales, en las que los términos teóricos juegan el papel de los estados intermedios en descripciones funcionales como las que hemos estado considerando hasta aquí. Como parte de su función explicativa, las teorías científicas típicamente introducen términos con significados nuevos. «Masa», por ejemplo, en la teoría newtoniana, es un término cuyo significado está emparentado con el significado cotidiano de ese término, o con el de «peso», pero no puede identificarse con él. La masa newtoniana de un objeto en la superficie de la Tierra o en la de la Luna es la misma, pero no su peso, en el sentido preteórico de esta noción. «Epiciclo» es un término teórico en la astronomía ptolemaica, «gen» lo es en la genética mendeliana e «inconsciente» lo es en la psicología freudiana. Lewis propone que el significado de todos estos términos está funcionalmente definido por el papel causal que se les asigna en las teorías respectivas, del mismo modo que el significado de la palabra «S_2», que el lector no conocía cuando se introdujo la descripción funcional F, está determinado por el papel causal que F asigna al estado S_2 en el proceso por ella definido. Así, la masa newtoniana de un objeto es aquella propiedad tal que: si ninguna fuerza actúa sobre el objeto, el producto de la misma

búho la descripción es sin duda implementada por un proceso neurológico, el psicólogo hará bien en tomar en consideración, al elaborar su propuesta, la información neurológica disponible. La propuesta final será independiente de la neurología del búho, en el sentido de que será realizable por procesos distintos a los neurológicos. Pero para ser una teoría psicológica verdadera de los búhos, ha de ser realizada en el búho por los procesos físicos que tienen lugar en estos animales.

8. La propuesta de Lewis se encuentra en Lewis (1972) y su fundamento teórico (su análisis de los términos teóricos), en Lewis (1983a).

por la velocidad del objeto permanece constante (primera ley del movimiento); si una fuerza actúa sobre el objeto, éste se acelerará en una cantidad directamente proporcional a la misma (segunda ley del movimiento); si el objeto ejerce una fuerza sobre otro, el segundo ejerce una fuerza sobre él igual en cantidad y dirección pero de sentido opuesto (tercera ley del movimiento), y tal que cualesquiera dos objetos ejercen uno sobre otro una fuerza atractiva proporcional al producto de la misma dividido por el cuadrado de la distancia entre ambos (ley de gravitación).

Por «teoría científica» entendemos generalmente un cuerpo de doctrina preciso, cuyo conocimiento requiere un aprendizaje especializado, etc. Mas las teorías científicas, en ese sentido usual, tienen predecesores que no se enseñan en las aulas universitarias, pero cuyo estatuto conceptual (aunque no epistemológico, predictivo o explicativo) es muy similar. Junto a la física «científica» está la física «popular», ese cuerpo de conocimientos que nos permite conducir una bicicleta, correr para coger el autobús o cazar una pelota al vuelo. Junto a la psicología científica está igualmente la «psicología popular». La psicología popular es la «teoría» que nos hace lanzarnos a cruzar una calle después de predecir primero que los conductores de los vehículos que se aproximan a gran velocidad hacia nosotros *percibirán* que la luz del semáforo está roja, y calcular después que, dado que *saben* que eso indica que deben detener su vehículo y *quieren evitar* las consecuencias de no hacerlo así, detendrán de hecho su vehículo. Pues bien, Lewis sostiene que los conceptos cotidianos de lo mental son conceptos teóricos, funcionalmente definidos a través de su papel en la psicología popular. (Esta versión del funcionalismo recoge también la de Putnam; pues la especificación de estados mentales que pueda hacer una psicología del futuro, en la forma de programas de ordenador, en que confía el funcionalista computacional, es también una descripción funcional en el sentido de Lewis.)

Por mor de la brevedad, en lo sucesivo consideraré de modo explícito únicamente el funcionalismo analítico que acabamos de presentar. Mas todo lo que se diga, tanto las consecuencias como las virtudes y las dificultades de esta variante de la concepción funcionalista de lo mental se aplica también, con modificaciones obvias, al funcionalismo computacional originalmente defendido por Putnam.

El cartesiano entiende la intencionalidad de lo mental sosteniendo que los estados mentales son estados de conocimiento consciente de ciertas cualidades fenoménicas concurrentes con el estado, cualidades independientes del «mundo externo». Es así que «roja» y «esfera» en «opino que hay una esfera roja ante mí» no refieren según él primariamente a características objetivas de las cosas, sino a cualidades fenoménicas. El conductista entiende la intencionalidad de un modo completamente distinto: según él, la intencionalidad es una manifestación del carácter disposicional de los estados mentales, del hecho de que los es-

tados mentales estén definidos en virtud de qué manifestaciones conductuales tendrían lugar en qué circunstancias. Los estados por relación a los cuales un estado mental está intencionalmente caracterizado son aquellos que sirven para definir las circunstancias externas que caracterizan la disposición conductual que constituya el estado. Es así que «roja» y «esfera» en «opino que hay una esfera roja ante mí» sí refieren según él a características objetivas de las cosas. El funcionalista, tanto computacional como analítico, está en esta cuestión del lado del conductista. La intencionalidad de lo mental consiste en el hecho de que los estados mentales son estados funcionalmente caracterizados: estados caracterizados por una teoría que especifica un modo de obtener ciertos *outputs*, dados ciertos otros *inputs*. Los estados con los que un estado mental está intencionalmente relacionado son aquellos que la teoría funcional invoca en la caracterización de los *inputs* y/o los *outputs*[9].

V. LA NECESIDAD DE SUPONER UN «LENGUAJE DEL PENSAMIENTO»

El lector no habrá dejado de reparar en la imprecisión de la propuesta funcionalista, tanto en su versión computacional como en su versión analítica. Mientras que la descripción funcional con la que introdujimos las nociones centrales de la propuesta funcionalista, F, era una descripción explícita y bien determinada, la propuesta funcionalista simplemente indica que los términos para lo mental son términos para estados internos en *ciertas* descripciones funcionales de las que sólo se dan algunas características. El funcionalismo computacional nos dice que se trata de las descripciones funcionales que la mejor teoría psicológica acabe proponiendo, el funcionalismo analítico que se trata de descripciones funcionales que de algún modo ya sabemos construir, utilizando para ello algunas de las leyes causales en las que figuran esos términos para lo mental que invocamos para explicar nuestra conducta y la de nuestros semejantes. Ambas propuestas funcionalistas nos dejan con la ansiedad que provocan las afirmaciones imprecisas.

Esta insatisfacción parece mayor en el caso del funcionalismo analítico. El funcionalista computacional sólo nos pide que esperemos unos años. La vaguedad de la propuesta del funcionalista analítico, sin embargo, no es de las que se resuelven con el paso del tiempo. Ya tenemos todos lo necesario para resolverla, en cierto sentido, pero el problema es que no sabemos por dónde empezar. Por ejemplo, ¿puede un ciego tener la opinión de que hay una esfera roja ante él? (No la *percepción*, sino la

9. Es natural pensar que la descripción de los outputs es esencial a la intencionalidad de los deseos, intenciones, etc., mientras que la de los inputs lo es a la de las opiniones, percepciones, saberes, etc. Los detalles varían con las diferentes concepciones del funcionalismo. Véase Block (1990) para una buena exposición breve, y Loar (1981, capítulo 4) para una detallada.

opinión.) Intuitivamente, parece que sí (el ciego puede haber aprendido a utilizar un espectrómetro). ¿Es esa opinión una del mismo tipo que la que puede tener una persona con visión normal? Si lo es, ¿qué leyes de la psicología popular, comunes al ciego y a la persona con visión normal, definen el término «opinión de que hay una esfera roja ante uno»? Si no lo es, ¿en qué medida se traduce esta diferencia en el significado de algunos términos mentales cuando se aplican a ciegos y cuando se aplican a individuos normales en diferencias en el significado de otros términos mentales (como «opinión de que hay un cubo ante uno» o «opinión de que hay un objeto físico ante uno» o incluso «opinión de que los triángulos rectángulos son pitagóricos»)? Dado que el funcionalista sostiene que el significado de los términos para lo mental está definido de un modo «holista», a través de su papel en una teoría que los invoca —la psicología popular—, parece que, si son diferentes teorías las que definen el significado de un término cuando se aplica a un individuo y cuando se aplica a otro, las diferencias se han de traducir en diferencias en el significado de todos los otros términos. Y podríamos seguir multiplicando de este modo la perplejidad.

Sin embargo, pese a tratarse de una propuesta imprecisa, el funcionalismo no es una propuesta vacía de contenido. Si lo fuese, no tendría las virtudes que en seguida enumeraremos, ni estaría sujeto a críticas como las que serán expuestas después. Y, por otra parte, el funcionalista puede decir, no sin cierto fundamento, que la vaguedad de que se le acusa es consustancial a todas las propuestas filosóficas. Ciertamente, es común a muchas de estas propuestas el tratarse de guías muy genéricas, y por tanto relativamente indeterminadas, acerca de la correcta perspectiva con que es conveniente mirar cierto ámbito. Es por consiguiente más la fertilidad al agilizar la investigación posterior, gracias a la eliminación de embrollos conceptuales, que la precisión la vara con que medirlas.

Existe, sin embargo, una diferencia de enorme calibre que no podemos pasar por alto entre nuestro paradigma de descripción funcional, F, y cualquier descripción funcional de la que razonablemente quepa esperar que defina los términos para lo mental. Esta diferencia tiene que ver con una característica de los estados mentales que se pone particularmente de relieve cuando consideramos los estados mentales que invocamos para dar cuenta de la competencia lingüística de los seres humanos, de su capacidad para usar el lenguaje. (Pero la característica en cuestión no es en absoluto privativa de esos estados mentales.)

Un modo de poner claramente de manifiesto la característica a que me refiero es insistir en que el número de estados mentales necesario para dar cuenta de la competencia lingüística de un ser humano normal es infinito, porque tal competencia lingüística es *productiva*. Se dice que una característica es productiva cuando, no por accidente, la tienen un número potencialmente infinito de objetos. Es razonable pensar que la

propiedad *ser una oración gramatical del castellano* es productiva: «la profesora de Daniel es inteligente» es gramatical, «la profesora de la profesora de Daniel es inteligente» también lo es, «la profesora de la profesora de la profesora de Daniel es inteligente» también lo es, y así sucesivamente, *ad infinitum*. Y también parece razonable decir que la competencia lingüística de un hablante normal es productiva, que un hablante normal *sabe* de un número potencialmente infinito de oraciones que son gramaticales.

Para defender esta afirmación de objeciones inmediatas es necesario indicar que nuestro uso común de «saber» y «opinar» otorga un carácter *potencial* a los estados a que nos referimos mediante ellos: no sólo *sabemos y opinamos* aquello que «tenemos presente», sino mucho más. Por ejemplo, si yo me dirijo hacia la silla con el fin de cambiarla de sitio es, entre otras cosas, porque *sé* u *opino* que la silla no se va a mover por sí sola. Si voy a comprar un billete para Nueva York a la agencia de viajes, es porque *sé* u *opino* que Nueva York está tan lejos que caminar hasta allí me llevaría demasiado tiempo y esfuerzo. Estas no son opiniones que yo tenga presentes, pero ciertamente son opiniones mías, pues formarían parte de una explicación exhaustiva de lo que hago. Es en este cotidiano sentido potencial de «opinar» y «saber» que es razonable decir que el número de opiniones y conocimientos que constituyen la competencia lingüística de un hablante normal es productiva.

Si esta tesis es correcta, encontramos aquí una asimetría radical entre nuestro paradigma de descripción funcional, F, y cualquier descripción funcional que pudiera servir para definir los términos para lo mental: una descripción funcional así habría de contener un número infinito de S_i, de expresiones para estados internos funcionalmente caracterizados. Si hemos de hacer plausible la propuesta funcionalista, hemos de dar alguna indicación de cómo puede haber una descripción funcional así. Y ciertamente, el funcionalista tiene algo que decir al respecto; su propuesta requiere introducir un aspecto consustancial al funcionalismo, la idea de un «lenguaje del pensamiento».

Conjeturar la productividad de la competencia lingüística es sólo un modo particularmente claro de hacer patente la necesidad de apelar al lenguaje del pensamiento. Dado que la productividad de la competencia lingüística resulta difícil de aceptar para algunos, será conveniente ver que hay un modo mucho menos comprometido de justificar su necesidad. Si, como trataremos de mostrar, dar cuenta de esta segunda característica en el marco funcionalista requiere también el lenguaje del pensamiento, entenderemos mejor por qué esta idea es ciertamente «consustancial» a la propuesta, cualesquiera que sean las dudas de algunos funcionalistas sobre el mismo.

La característica en cuestión es la *sistematicidad* de la competencia lingüística. Compárese el conocimiento que tiene un hablante competente

del castellano de la gramaticalidad y del significado de (1) «la profesora de Daniel es inteligente» con el conocimiento que tiene un individuo que no conoce el castellano y ha aprendido que esa oración es gramatical, y cuál es su significado, al modo en que los turistas aprenden la corrección gramatical y el significado de una lista más o menos grande de las oraciones del lenguaje del país que se disponen a visitar. Un síntoma de la diferencia es que, en el primer caso, pero no en el segundo, esa competencia basta para conocer también tanto la corrección gramatical como el significado de (2) «la profesora de la profesora de Daniel es inteligente». Se dice que una propiedad es sistemática cuando el que se aplique o no a un objeto en su dominio depende necesariamente de la estructura del objeto: depende de que el objeto esté compuesto de otros, con ciertas otras propiedades y estando en ciertas relaciones entre sí. La competencia lingüística de un hablante normal es sistemática porque, por ejemplo, su conocimiento de la corrección gramatical y del significado de (1) consiste en su conocimiento de las categorías gramaticales de las expresiones componentes, de sus significados, de los modos de combinación de las categorías que dan lugar a oraciones correctas, y de los significados de las expresiones así combinadas. Es esto lo que explica el síntoma de la sistematicidad antes mencionado: a saber, que el individuo sea también capaz de entender (2), en que las mismas palabras están combinadas «más veces» con arreglo a idénticos modos de combinación. Por contra, dado que el conocimiento que el turista tiene del significado y la corrección de (1) no es sistemático, sino que atribuye esas características «como un todo» a (1), ese conocimiento no le basta para entender también (2).

No sólo los estados mentales que constituyen la competencia lingüística son sistemáticos; éste es sólo un caso particularmente obvio. En rigor, cabe pensar (y existen razones profundas para ello) que la sistematicidad es una marca característica de lo mental. El conocimiento que tiene un búho de la ubicación en el espacio externo de ciertos sonidos trasciende los sonidos de que de hecho ha tenido experiencia a lo largo de su existencia: si le presentamos sonidos que nunca antes había oído, será capaz de ubicarlos con igual precisión en el espacio y seguirlos a través del tiempo. Este es un síntoma de la sistematicidad de una propiedad, en este caso una propiedad mental: la percepción de sonidos en el espacio.

Pues bien, no sólo la discutible productividad, sino también la indiscutida sistematicidad ofrece un contraste fundamental entre descripciones funcionales como nuestro paradigma, F, y cualquier descripción funcional de la que razonablemente quepa pensar que define los términos para lo mental. Dar cuenta de la sistematicidad de un conjunto de estados mentales (como los que constituyen la competencia lingüística de un hablante normal) requiere que la descripción especifique algunos de los estados S_i (por ejemplo, el estar en el estado de saber que una cierta ora-

ción es gramatical, o que tiene un cierto significado) como el resultado necesario de estar en estados correspondientes a las «partes» pertinentes. Esto mismo se requiere también para dar cuenta de la productividad de un conjunto de estados mentales —los que constituyen la competencia lingüística otra vez: de hecho, es razonable pensar que la productividad de un conjunto de estados mentales, cuando se trata de los estados mentales de un ser humano y no de los de Dios, implica su sistematicidad—. Pero su sistematicidad no implica ciertamente su productividad: hay sin duda lenguajes artificiales finitos cuyo conocimientono es menos sistemático que el de los lenguajes naturales. He aquí la justificación de que aseverar la sistematicidad de una cierta competencia sea hacer una afirmación más débil, y por tanto menos discutible, que aseverar la productividad de esa competencia [10].

El funcionalista está en la obligación de decirnos cómo su propuesta puede acomodar las exigencias impuestas por la sistematicidad de muchos estados mentales (quizás de todos). Hasta ahora contamos con un ejemplo (la descripción funcional F) y una atrevida generalización a partir del mismo; pero el ejemplo no nos da ninguna idea de cómo acomodar los nuevos requisitos. Es en este contexto en el que el funcionalista recurre a la hipótesis del «lenguaje del pensamiento». En realidad, la hipótesis la sugiere ya el paradigma de descripción funcional que el funcionalista suele tener en mente, que no es ni mucho menos algo tan humilde como F, sino más bien un programa de ordenador capaz de producir tareas complejas, como operaciones aritméticas o proceso de textos.

Un modo de explicar la sistematicidad de un conjunto de estados funcionalmente caracterizados es invocar en la caracterización funcional un mecanismo que produce los estados constitutivos de la capacidad sistemática como el resultado de un proceso sistemático. Por ejemplo, podemos decir que el conocimiento de la gramaticalidad de (1) es el resultado de un proceso que involucra (i) estados que representan que la categoría de «la» es *Det*, la de «profesora» *NC*, la de «de» *Prep*, la de «Daniel» *N*, la de «es» *VC*, la de «inteligente» *PN*; (ii) estados que representan que un *Det* seguido de un *NC* es un *N*, que un *Prep* seguido de un *N* es un *Ad*, y que un *N* seguido de un *Ad* es un *N*, y (iii) estados que

10. Véase Fodor (1987), para la mejor defensa del «lenguaje del pensamiento», así como para las nociones de productividad y sistematicidad. Históricamente, las críticas de Noam Chomsky al conductismo skinneriano y sus propias ideas sobre la competencia lingüística fueron una fuente de inspiración crucial para el funcionalismo. No es muy arriesgado pensar que la conjetura chomskyana de que la competencia lingüística de un hablante normal estaría constituida por una formulación interna de algo sustancialmente similar a una teoría lingüística (de la que el hablante no tiene conocimiento consciente) fue un modelo fundamental tanto para sugerir la propuesta funcionalista como la hipótesis de un «lenguaje del pensamiento». Éste sería así el «lenguaje» en que la «teoría» constitutiva de la competencia estaría «formulada».

representan que un *N* seguido de un *VC* y de un *Ad* es una oración gramatical. El lector puede comprobar que la existencia de estos estados determina la existencia «potencial» de un estado que representa la gramaticalidad de (1) y también la de uno que determina la gramaticalidad de (2)[11]. Lo que hemos hecho es presentar los estados que constituyen el conocimiento de la gramaticalidad de (1) y el de la de (2) como el resultado de ciertos subprocesos. La relación sistemática que observamos entre el conocimiento de la gramaticalidad de (1) y el de la gramaticalidad de (2) la explicamos por el hecho de que los subprocesos implicados en ambos casos tienen partes en común.

El carácter intencional o representacional de los estados mentales se explica en la concepción funcionalista por el hecho de que la caracterización de un estado funcional requiere especificar ciertos *inputs* y ciertos *outputs*, además de un proceso que los conecta: los estados con los que un estado mental está intencionalmente relacionado (aquellos que «representa») son los implicados en la descripción de los *inputs* y/o los *outputs* pertinentes. El lector apreciará que los estados intermedios invocados en los subprocesos a que apelamos en una explicación de la sistematicidad como la ejemplificada en el párrafo anterior, *son ellos mismos representacionales*, en el sentido indicado: los subprocesos invocan subestados que «representan», esto es, subrepresentaciones. Es así como la explicación de la sistematicidad a que acabamos de apelar supone un «lenguaje del pensamiento». Las «representaciones» que, incorporadas al modo ejemplificado en descripciones funcionales, darían cuenta de la sistematicidad, son representaciones *lingüísticas* en un sentido muy importante. Primero, tienen sintaxis: las representaciones postuladas en los estados en (i) tienen un «sujeto», para la expresión, y un «predicado», para la categoría; las postuladas en (ii) tienen estructura condicional. Segundo, tienen *interpretación semántica*: los sujetos de las representaciones en (i) significan diferentes palabras del castellano, y sus predicados diferentes categorías sintácticas (diferentes conjuntos de palabras del castellano, si se quiere).

La propuesta del «lenguaje del pensamiento» consiste en que el ejemplificado no es sólo un modo entre otros de explicar la sistematicidad de los estados mentales de seres como nosotros (y también los de los búhos), sino el único empíricamente plausible. Para evaluar propiamente la propuesta, empero, es muy importante apreciar que las representaciones mismas son parte de la descripción funcional, y que están ellas mismas funcionalmente caracterizadas. Es decir, la propuesta sólo exige que

11. Junto con otros que determinan la gramaticalidad de oraciones que lo son en castellano, y también la de algunas que, desgraciadamente, no lo son: lo que muestra que la descripción funcional correcta de la competencia lingüística de un hablante del castellano ha de ser algo más complicada y cuidadosa.

cualquier objeto que realice una descripción funcional así tenga componentes que *funcionen* como las representaciones en cuestión, que tengan ese papel causal. Decir que una descripción funcional que incorpore (i), (ii) y (iii) ha de ser parte de la caracterización funcional de mi competencia lingüística no es decir que, si se examinara mi cerebro mientras determino que (1) es gramatical, se encontrarían combinaciones de neuronas con la apariencia de oraciones del lenguaje natural, escrito o hablado, de la forma: «la» es un *Det*, etc. Esto es simplemente absurdo. Sólo es decir que, en cualquier descripción neurológica correcta de los dos procesos que hacen que yo reconozca primero (1) como gramatical y después (2) como gramatical, se deben discernir partes comunes, correspondientes al reconocimiento de «la», a la asignación de la categoría «Det» a «la», etc., operando en el proceso en la secuencia causal descrita por la descripción funcional. Pero esas partes no tienen por qué estar combinadas como lo están las oraciones del lenguaje natural, en concatenación espacial o temporal, por ejemplo. Lo más probable es que no lo estén así. (En los ordenadores tradicionales, que operan con «representaciones» y un «lenguaje del pensamiento» en el sentido funcional que se acaba de exponer, *no* lo están.)

La postulación de un «lenguaje del pensamiento» como el único modo plausible de explicar la sistematicidad de nuestras capacidades mentales ha sido contestada empíricamente por los partidarios del conexionismo; algunos funcionalistas se resisten independientemente a ella, por considerar que el funcionalismo debería ser una hipótesis mínima, no comprometida con hipótesis según ellos tan fuertes sobre la estructura interna de la mente [12]. En mi opinión, una actitud tan mínima sólo es compatible con el conductismo, no con el funcionalismo (aunque para algunos el funcionalismo es una variante del conductismo con muy poco o ningún contenido adicional). Pero el funcionalismo, tal como aquí se ha expuesto, es una genuina alternativa al conductismo; que los estados mentales son estados funcionalmente descritos implica que cuando decimos que alguien está en un cierto estado mental estamos adquiriendo importantes compromisos sobre la estructura causal interna de ese estado, aun si más abstractos que los que adquiriríamos si lo describiéramos neurológica o químicamente. Cuanto más complejas sean las correlaciones entre *inputs* y *outputs* mediante en términos de cuyas explicaciones definimos funcionalmente estados mentales, más elevado nuestro compromiso sobre la estructura interna que habremos de postular y asumir como supuesto de nuestra definición. Una concepción funcionalista de los estados mentales de individuos capaces de conductas tan complicadas como los seres humanos ha de adquirir, necesariamen-

12. Daniel Dennett, quien por otro lado se considera funcionalista, es un conocido adversario de la tesis. Véase, por ejemplo, «Brain Writing and Mind Reading», en Dennett (1978).

te, un alto compromiso ontológico sobre la naturaleza de los estados internos postulados.

Por otra parte, si uno no se deja engañar por la falsa analogía con el lenguaje natural sugerida por la etiqueta «hipótesis del lenguaje del pensamiento», y se atiene al contenido estrictamente funcionalista de la hipótesis, el hecho de la sistematicidad de lo mental, especialmente allí donde es más manifiesto (en el caso de la competencia lingüística), le confiere a la hipótesis una plausibilidad muy grande. Al menos, en el supuesto funcionalista de que el concepto de un estado mental es el concepto de un estado funcionalmente caracterizado. Tanta que, en mi opinión, las consideraciones de los conexionistas están aún muy lejos de ponerla en duda (y más bien hacen dudar de la viabilidad del programa conexionista, en la medida en que el programa sea realmente incompatible con la hipótesis del lenguaje del pensamiento).

VI. VIRTUDES Y DEFECTOS DE LA PROPUESTA FUNCIONALISTA

Expondremos finalmente las virtudes de la propuesta funcionalista, en relación con las dos antes consideradas (la cartesiana y la conductista), y las principales dificultades con que tropieza. Estas últimas serán discutidas con mayor detalle en capítulos posteriores.

El mayor mérito del funcionalismo consiste en reunir las virtudes del conductismo sin los defectos de esta propuesta. Para apreciar del modo más profundo por qué esto es así, hemos de ver que la tesis funcionalista no es más que la misma tesis conductista (los estados mentales son disposiciones a la conducta) entendida bajo el supuesto de una concepción distinta de la noción de *disposición*. El conductista concibe una disposición como definida por los condicionales subjuntivos que enuncian sus manifestaciones; la solubilidad de un objeto consiste en que si se le pusiera en agua se disolvería. Y estos condicionales subjuntivos los interpreta en términos sólo epistemológicos, carentes de compromiso ontológico sustancial: de nuevo, en afortunada expresión de Ryle, como «autorizaciones para la inferencia»: la autorización para inferir que un objeto se disolverá, cuando disponemos de la información de que el objeto ha sido puesto en agua. Atribuir una disposición a un objeto, en esta concepción, no es hablar de la explicación de una regularidad observable, sino sólo de la regularidad observable misma: hablar de lo que podemos esperar observar dadas ciertas circunstancias observables.

Según esto, atribuir a un individuo la opinión de que hay una esfera roja ante él es autorizarnos a inferir que de su boca saldrá el sonido «sí» si disponemos de la información de que se le ha preguntado si hay una esfera roja ante él, etc. Es esta concepción de las disposiciones la que provoca las dificultades del conductista; pues el hecho de que tales autori-

zaciones para la inferencia parezcan ser siempre relativas a información sobre la existencia de otros estados mentales (el holismo de lo mental) nos hace sospechar de la existencia de una ineliminable circularidad en la tesis conductista. Y el caso de los super-superespartanos contrasta la intuición de que los estados mentales son genuinamente «internos» con esta concepción de las disposiciones; pues se trata de un ejemplo de existencia de estados mentales que no va acompañada por ninguna autorización a la inferencia, supuesta cierta información, de que cierto comportamiento observable tendrá lugar.

Es esta concepción «verificacionista» de las disposiciones la que provoca las dificultades del conductismo, y no la tesis misma de que los estados mentales son disposiciones. Esto puede comprobarse abandonando la concepción verificacionista de las disposiciones en favor de una «realista». De acuerdo con esta nueva concepción, una disposición es un estado interno definido a través de una descripción funcional; la descripción funcional está compuesta de esos condicionales con que el verificacionista identifica la disposición. Por ejemplo, la solubilidad es un estado caracterizado por tratarse del estado interno de un objeto que causa que el objeto se disuelva cuando se le pone en agua. (Decir «causa» es ya, implícitamente, hablar en subjuntivo: si el estado *e* —por ejemplo, una cierta composición química— causa que el objeto se disuelva cuando se le introduce de hecho en agua, ese mismo estado *causaría* que se disolviera si se le pusiera en agua.) Los condicionales definen funcionalmente el estado interno, pero la solubilidad es ese estado interno mismo, y no meramente una autorización a inferir ciertas cosas supuestas ciertas otras.

Vistas así las disposiciones, podemos seguir sosteniendo que los estados mentales son disposiciones sin caer en las dificultades del conductismo. En primer lugar, el holismo no presenta ahora ninguna dificultad. Como hemos visto mediante el ejemplo de la descripción F, una descripción funcional puede definir simultáneamente muchos estados internos. Los estados mentales, vistos como disposiciones «realistas», muy bien pueden estar causalmente interrelacionados, y ser por consiguiente necesariamente interdefinibles en una concepción funcional. Incluso cabe decir que es la teoría psicológica completa (científica o popular) la que define como un todo cada una de esas disposiciones que son los estados mentales, a través del papel causal respectivo que la teoría les asigna. En segundo lugar, los estados mentales, así concebidos, son genuinos estados internos. No existe dificultad en pensar ciertos individuos que están en el estado interno constitutivo del dolor (definido funcionalmente como el estado que causa gemidos cuando a uno le sacan una muela sin anestesia *y no desea con un deseo grandemente intenso evitar gemir*, etc.), aunque en esos individuos ese estado no causa las mismas manifestaciones que en otros, que también están en él, sencillamente

porque los individuos en cuestión están también, a diferencia de los otros, en otro estado interno (a saber, el deseo grandemente intenso de ahogar los gemidos) que conspira (causalmente hablando) para impedir que el primero tenga sus resultados característicos [13].

Es así que el funcionalismo recoge los méritos del conductismo evitando sus dificultades. El funcionalismo está caracterizado por la misma «inocencia semántica» que el conductismo, y por tanto evita el problema del escepticismo sobre la mente de los demás. Cuando decimos que alguien tiene una opinión de que hay una esfera roja ante él, estamos empleando «rojo» con el mismo significado que cuando decimos que hay una esfera roja ante uno, aunque con una función distinta: en lugar de usar la expresión «rojo» para atribuir una cierta propiedad a un objeto, la usamos para especificar funcionalmente un estado caracterizado en parte por ser uno causado por la presencia de un objeto con la propiedad a que nos referimos con «rojo». Por otra parte, en el supuesto de que pueda distinguir, de entre todas las disposiciones, aquellas que determinan una distinción entre manifestaciones correctas e incorrectas (y no meramente manifestaciones probables o improbables), el funcionalismo no está sujeto a las objeciones basadas en la cuestión de la normatividad de lo mental que Wittgenstein dirigió contra la concepción cartesiana de la mente en su famoso argumento contra la posibilidad de un lenguaje privado. Esta distinción entre disposiciones normativas y no normativas puede hacerse según las líneas «sociales» del propio Wittgenstein, o, en mi opinión de un modo más plausible, introduciendo elementos teleológicos prestados de la biología en la noción de «función» (aquí expuesta en términos meramente causales).

Pasemos para concluir a la enumeración de las dificultades, comenzando, a modo de transición, por la cuestión de la eficacia causal, que está, tanto para el conductismo como para el funcionalismo, curiosamente a medio camino entre las virtudes y los defectos. Existe a este respecto un cisma en las filas funcionalistas, el que separa a los funcionalistas «de primer orden» (Lewis es el ejemplo más notorio) y a los funcionalistas «de segundo orden» (Putnam, y de hecho casi todos los demás).

Volvamos al ejemplo más simple de la solubilidad, pues el hecho de que las disposiciones constitutivas de los estados mentales posean defi-

13. Adversarios del funcionalismo como Chomsky y Searle tienden a enfatizar la similitud entre conductismo y funcionalismo que aquí estamos poniendo de relieve: conductismo y funcionalismo pueden verse ambos como la tesis de que el concepto de un estado mental es el concepto de una disposición a la conducta. Pero sería un error pasar por alto la fundamental diferencia en el modo de entender las disposiciones que esta tesis común encubre. Es consustancial al conductismo la concepción verificacionista de las disposiciones, pues el conductismo es una elaboración del rechazo a una concepción de la mente (la cartesiana) que hace de la introspección el acceso epistémico privilegiado a la mente. El conductista, por contra, quiere sustituir la concepción cartesiana de la mente por una en que nuestros estados mentales son manifiestos a los demás.

niciones mucho más complejas, en que se definen muchos estados a la vez, y el hecho de que sean disposiciones normativas —si esto se interpreta en términos teleológicos— no afecta a la cuestión, sólo la complica. Supongamos que el estado constitutivo de la solubilidad en un determinado objeto (el estado que explica causalmente que el objeto se disuelva si se le pone en agua) es un cierto estado químico, Q. ¿Cabe identificar la solubilidad con Q? Eliminemos una potencial ambigüedad contenida en esta pregunta antes de responderla. Por supuesto, las expresiones «Q» y «solubilidad» tienen significados distintos, expresan conceptos distintos. Lo que nos preguntamos es si cabe efectuar una *identificación teórica*, como aquella que, en virtud de la reducción de la termodinámica clásica a la mecánica estadística, llevamos a cabo cuando decimos que el calor es la energía cinética media de las partículas. Como también los predicados «calor» y «energía cinética media» expresan, sin duda, diferentes conceptos, podemos decir que en una identificación teórica se identifican las *propiedades* indicadas por las expresiones a ambos lados de la expresión para la identidad. Supondremos que las propiedades, a diferencia de los conceptos, son entidades «objetivas» que tienen que ver con los poderes causales de las cosas y la explicación de los procesos en que intervienen.

La pregunta sobre la posibilidad de identificar la solubilidad con Q, pues, concierne a la identidad de *la propiedad* de la solubilidad con la propiedad química Q: identificar los conceptos sería manifiestamente incorrecto. El funcionalista de primer orden responde afirmativamente a la misma: una propiedad funcionalmente caracterizada es, sencillamente, la propiedad física que la realiza. Supuesta esta versión del funcionalismo, no hay ningún conflicto entre la tesis de la eficacia causal de la mente y la tesis de la completitud causal del mundo físico. Cuando retrotraemos causalmente a partir del movimiento de mi brazo, sólo encontramos estados neurológicos; pero uno de estos estados neurológicos, es de esperar, *realiza* mi deseo (funcionalmente entendido) de coger algo rojo (y por tanto es idéntico a él), y otro mi creencia de que hay una esfera roja ante mí. Por consiguiente, los estados mentales son causalmente eficaces, tanto como los estados neurológicos que los realizan, con los que se identifican.

El funcionalista de segundo orden sostiene, por su parte, que esta idílica representación está muy alejada de la realidad. Los estados funcionales medianamente interesantes son *múltiplemente realizables*: en diferentes objetos, los estados físicos que los realizan son físicamente distintos. Dos ordenadores, por ejemplo, pueden muy bien estar ambos en el estado que realiza un mismo estado funcionalmente caracterizado por un mismo programa (ambos están ejecutando Word Perfect, y ambos están ejecutando la instrucción de copiar una misma palabra en documentos idénticos hasta la última coma), y, sin embargo, el estado físico

en el que están puede ser muy distinto. Aunque es menos plausible, lo mismo puede ocurrir con la solubilidad. Supongamos que así ocurre: tanto un terrón de A como uno de B se disolverían si se pusieran en agua, pero la explicación de que así ocurra apela a características físicas muy distintas en el caso de A y en el de B. Esto significa que muchos procesos causales que desde el punto de vista macroscópico son todos ellos manifestaciones de la solubilidad, desde el punto de vista de los estados microfísicos que los realizan son procesos muy distintos. Los procesos que desde el punto de vista computacional son todos ellos idénticos, pueden ser sin embargo físicamente procesos muy distintos entre sí. Dada esta situación, ¿tiene sentido identificar la solubilidad, o la propiedad computacional mencionada arriba, con la propiedad física que la realiza en un caso particular?

Hay buenas razones para concluir que no. Las identificaciones teóricas tienen sentido cuando los tipos de proceso causal en que intervienen las propiedades reducidas tienen un reflejo suficientemente fiel en procesos en que intervienen las propiedades reductoras. Esto es lo que ocurre en el caso antes mencionado de la reducción de la termodinámica a la mecánica estadística. Pero nada así es el caso en una situación de múltiple realizabilidad. Procesos que desde el punto de vista funcional son el mismo (en todos ellos interviene el estado de copiar una cierta palabra en un documento de cierto tipo, por ejemplo), son sin embargo muy distintos entre sí cuando se examinan microfísicamente. La múltiple realizabilidad es una razón para considerar a las propiedades funcionales propiedades en sí mismas, autónomas, caracterizadas por intervenir en ciertos procesos causales macroscópicos (aquellos que las definen funcionalmente, y que se atomizan en miríadas de procesos distintos cuando se examinan microfísicamente). El funcionalismo de segundo orden recoge esta idea, identificando las propiedades funcionales múltiplemente realizadas con propiedades de segundo orden, definidas haciendo mención genérica de propiedades de primer orden (presumiblemente físicas): la propiedad de tener una propiedad de primer orden u otra que cumple un cierto papel causal. La solubilidad (supuesta su múltiple realización) sería la propiedad que tiene un objeto de tener una propiedad física u otra que causa que el objeto se disuelva cuando se le pone en agua [14].

Frente al funcionalismo de primer orden, el funcionalista de segundo orden argumenta que su funcionalismo concede un papel autónomo a la

14. Putnam comenzó defendiendo el funcionalismo de primer orden en los primeros artículos en que expuso la concepción funcionalista. Véase, por ejemplo, «Minds and Machines» (originalmente publicado en 1960), en H. Putnam, *Philosophical Papers*, vol. 2, Cambridge (1975). Putnam (1975b), defiende ya el funcionalismo de segundo orden, apelando esencialmente a la «múltiple realizabilidad» de los estados mentales. Para argumentos similares, véase Block y Fodor (1972) y Fodor (1974). Una defensa del funcionalismo de primer orden puede verse en Lewis (1983b).

psicología. El funcionalista de primer orden recoge ciertamente la eficacia causal de las propiedades mentales, pero a costa de identificarlas con propiedades físicas. Además, las propiedades mentales, pensadas como funcionalmente caracterizadas, parecen ser múltiplemente realizadas: cuenta en favor de ello la intuición de que un marciano podría muy bien tener una opinión de que hay una esfera roja ante él, aun siendo físicamente muy distinto a un ser humano; y más aún el hecho, comprobado por la neurología, de la plasticidad neurológica del cerebro humano. (En individuos en los que un área del cerebro, que en seres normales parece realizar una cierta función mental, ha resultado lesionada, otra área se especializa en realizar esa función.) Si eso fuera así, las regularidades causales que la psicología trata de formular, y en cuya formulación figuran las propiedades mentales, serían invisibles al nivel físico: al nivel físico lo que psicológicamente es el mismo proceso, y por tanto involucra la misma propiedad, serían muchos procesos muy distintos entre sí, involucrando propiedades totalmente disímiles. Para el funcionalista de segundo orden, nada de esto atentaría contra la existencia de propiedades específicamente psicológicas, concebidas como propiedades funcionales de segundo orden: el estado consistente en estar en un estado físico u otro caracterizado por realizar un cierto poder causal.

El funcionalismo de segundo orden es la versión del funcionalismo más compatible con los hechos; y la ironía reside en que una teoría elaborada en buena medida para soslayar las dificultades cartesianas con la eficacia causal de la mente se enfrenta ahora, como el funcionalista de primer orden hace notar, con dificultades similares. Pues consideremos dos individuos que mueven sus brazos de maneras similares para alcanzar esferas rojas similares similarmente situadas ante ellos. El funcionalista de segundo orden sostiene que la causa es en ambos casos la misma, una causa mental: la propiedad común a ambos de tener la creencia de que hay una esfera roja ante ellos junto con el deseo de coger algo rojo. Supongamos además que los estados neurológicos que realizan esa propiedad mental son distintos en ambos casos. Lo que el funcionalista de primer orden enfatiza es que los estados neurológicos respectivos, distintos en ambos casos, ofrecen una explicación causal completa y satisfactoria de los respectivos movimientos de los brazos. En la medida en que la presunta causa mental sea distinta de ellos, no juega ningún papel en la explicación causal. Pretender lo contrario sería introducir, en cada caso, una causa adicional a otra que por sí misma es enteramente explicativa: sería, pues, postular la existencia sistemática de una sobredeterminación causal, sin ninguna justificación para ella. Más le vale al funcionalista de segundo orden considerar a sus propiedades causalmente ociosas, epifenoménicas. El funcionalismo, pues, tiene aún dificultades con la cuestión de la eficacia causal de la mente. El funcionalismo de primer

orden parece hacer a la mente redundante, mientras que el funcionalismo de segundo orden la hace ociosa [15].

Otras dificultades del funcionalismo (comunes también al conductismo) se originan en tropiezos con las firmes intuiciones que alimentan la concepción cartesiana de la mente. La primera es el carácter *intrínseco* de lo mental. Intuitivamente, que yo tenga la opinión de que hay una esfera roja ante mí no depende en absoluto de que yo esté en una cierta relación causal con nada externo: yo podría tener una opinión exactamente con ese contenido, incluso aunque no hubiese nada externo. Esta intuición se puede explotar para construir nuevos argumentos contra el funcionalismo. Sin entrar en los detalles, en la versión del funcionalismo que tiene todas las virtudes mencionadas arriba, el funcionalismo teleológico, los estados mentales son extrínsecos. Tener una opinión de que hay una esfera roja ante uno es estar en un estado que ha sido seleccionado, mediante la selección natural y quizás también mediante el aprendizaje a través de la experiencia del individuo, para «activarse» como consecuencia de la presencia en el entorno de una cierta situación objetiva, caracterizada por la presencia de un objeto capaz de reflejar luz en una cierta banda de longitudes de onda. Ahora bien, parece perfectamente posible concebir un individuo «intrínsecamente» exactamente como yo, ahora, en que opino que hay una esfera roja ante mí; tan exacto, que es no sólo idéntico en cuanto a las «cualidades» sensoriales que se le presentan, sino que neurológicamente es totalmente idéntico a mí; y, sin embargo, que el estado físico que en mí ha sido seleccionado para activarse como consecuencia de la presencia en el entorno de un objeto capaz de reflejar luz en una cierta banda de longitudes de onda, y en el que yo estoy ahora, *ese mismo estado físico*, en él ha sido seleccionado para activarse como consecuencia de la presencia en el entorno de un objeto capaz de reflejar luz en *otra* banda de longitudes de onda (digamos la que en mí produce la cualidad que yo llamo «verde»); y que en su presencia ahora hay de hecho un objeto reflejando luz en esa *otra* banda «verde», mientras que en la mía hay uno reflejando luz en la banda «roja». Esta situación, una típica situación de las de «inversión de espectro» invocadas en la filosofía tradicional, parece, como se dijo, perfectamente concebible; y nuestras intuiciones apuntan además a que en un sentido muy importante de «estado mental», los dos individuos en ella comparten el mismo estado mental. Sin embargo, es claro que, si los estados mentales están funcionalmente caracterizados, los dos individuos están en distinto estado mental. Algunos funcionalistas han tratado de acomodar algunas de las intuiciones que sustentan experimentos mentales como el anterior

15. El argumento expuesto contra el funcionalismo de segundo orden lo invoca usualmente David Lewis. Véase, por ejemplo, Lewis (1986). También puede encontrarse en varios artículos de Jaegwon Kim, recogidos en Kim (1993).

tratando de introducir una noción más «intrínseca» pero igualmente funcional de contenido representacional, al que han bautizado como «contenido *narrow*», «contenido estrecho». Se trataría de un contenido determinado funcionalmente, y por tanto causalmente, pero sin apelar más que a estados «de la piel para adentro». Sin embargo, ninguna de las caracterizaciones hasta el momento ha sido muy precisa; y existen dudas más que razonables respecto de que cualquiera de ellas pueda recoger una noción intuivamente plausible de *contenido intencional*[16].

El gran tema pendiente, sin embargo, es el de la consciencia (la cuestión ya estaba implícita en la discusión anterior, bajo el término «intrínseco»). Parece muy difícil que la maquinaria funcionalista, con su apelación para la definición de lo mental a una ingente suma de relaciones causales, científicamente establecidas o *folk*, pueda acomodar las intuiciones sobre ese peculiar tipo de conocimiento de sí, con sus características de inmediatez y certidumbre, que es constitutivo de lo que paradigmáticamente llamamos «estados conscientes». Y la alternativa eliminacionista, por ingeniosas que sean las consideraciones tendentes a poner en duda la coherencia de la noción de consciencia, parece estar sencillamente fuera de lugar[17]. El neocartesiano no tiene ninguna razón para cantar victoria ante esta dificultad, sin embargo, por cuanto un juez imparcial sin duda manifestaría que sus esfuerzos no han dado mejor fruto. Formular una explicación satisfactoria del concepto de consciencia, dentro o fuera del marco funcionalista, es la tarea a la vez inaplazable e ingrata para esa aspiración a saber de qué se habla que, desde Sócrates, anima la empresa filosófica.

VII. RESUMEN Y CONCLUSIONES

En este capítulo hemos presentado la propuesta funcionalista, distinguiéndola dialécticamente de sus rivales, el conductismo y la concepción cartesiana de la mente. Frente a la segunda, y como en la primera, en la concepción funcionalista la consciencia no se considera una característica esencial de la mente; por el contrario, lo esencial de los estados mentales es su carácter disposicional. Frente al conductismo, por otro lado, en la concepción funcionalista los estados mentales son genuinos estados internos, que median causalmente en la producción de la conducta observable en las entidades a que se les atribuyen, y no son meros sumarios de nuestras expectativas sobre su comportamiento observable en cir-

16. Véase Fodor, *Psychosemantics*, MIT Press, Cambridge, Mass., 1987, capítulo 2, para una elaboración de la idea de «contenido estrecho». Véase García-Carpintero (1993 y 1994), para réplicas a argumentos como el considerado que no requieren introducir tal noción.

17. Los argumentos eliminacionistas más ingeniosos se deben a Daniel Dennett. Véase, por ejemplo, Dennett (1991).

cunstancias observables. Además, hemos distinguido diversas variedades de la propuesta funcionalista. Según que la teoría que define funcionalmente los estados mentales sea una teoría científica (el producto final de los esfuerzos de los psicólogos cognitivos) o la «psicología popular» común a casi todos los seres humanos (quizás con la excepción de los autistas), y en la que los «expertos» serían individuos como Proust, los sacerdotes católicos y los consejeros matrimoniales, tenemos el «funcionalismo computacional» o el «funcionalismo analítico». Cada una de estas propuestas, a su vez, puede ser dividida ulteriormente, según que se suponga que los estados funcionalmente caracterizados se identifican con los estados físicos que los realizan («funcionalismo de primer orden») o que se sostenga más bien que los estados funcionales son ellos mismos estados autónomos, el estado en que se está cuando se está en uno u otro de los diversos estados físicos que realizan o pueden realizar una cierta caracterización funcional («funcionalismo de segundo orden»). Por último, hemos mostrado cómo el funcionalismo se ve comprometido con la idea de un «lenguaje del pensamiento».

En lo que respecta a cuestiones evaluativas, hemos mostrado que el funcionalismo, en alguna de sus versiones, es una propuesta conceptualmente atractiva, capaz de arrojar luz sobre muchos de los tradicionales problemas filosóficos que configuran el ámbito de la filosofía de la mente. Sin embargo, no hemos ocultado que no es ni mucho menos la funcionalista una propuesta que permita esperar el fin de los debates en ese ámbito, como lo ponen de manifiesto los diversos problemas que hemos señalado: la cuestión de la eficacia causal de los estados funcionalmente caracterizados, la naturaleza intuitivamente intrínseca de los estados mentales y, por encima de todo, la naturaleza de la consciencia.

BIBLIOGRAFÍA

Block, N. (1990), «The Computer Model of the Mind», en D. Osherson y E. Smith (eds.), *Thinking*, MIT Press, Cambridge, Mass.

Block, N., y Fodor, J. (1972), «What Psychological States Are Not»: *Philosophical Review* lxxxi, 159-181.

Carnap, R. (1932-1933), «Psychologie in physikalischer Sprache»: *Erkenntnis, 3* (1932-1933), 107-142. V.e. en A. Ayer (ed.), *El positivismo lógico*, FCE, México, 1965.

Cirera, R. (1990), *Carnap i el Cercle de Viena*, Anthropos, Barcelona.

Dennett, D. (1978), *Brainstorms*, MIT Press, Cambridge, Mass.

Dennett, D. (1991), *Consciousness Explained*, Little, Brown & Company, Boston.

Fodor, J. (1974), «Special Sciences, or The Disunity of Science as a Working Hypothesis»: *Synthese, 28*, 97-115.

Fodor, J. (1987), «Why There Still Has to Be a Language of Thought», en Id., *Psychosemantics*, MIT Press, Cambridge, Mass.

García-Carpintero, M. (1993), «The Supervenience of Mental Content»: *Proceedings of the Aristotelian Society*, 119-135.
García-Carpintero, M. (1994), «Dretske on the Causal Efficacy of Meaning»: *Mind and Language*.
Geach, P. (1992), *Mental Acts*, Thoemmes Press, Bristol [1]1957.
Kim, J. (1993), *Supervenience and Mind*, Cambridge University Press, Cambridge.
Lewis, D. (1972), «Psychophysical and Theoretical Identifications»: *Australasian Journal of Philosophy*, *50*, 249-258.
Lewis, D. (1983a), «How to Define Theoretical Terms», en Id., *Philosophical Papers,* vol. 1, OUP, Oxford.
Lewis, D. (1983b), «Mad Pain and Martial Pain», en Id., *Philosophical Papers,* vol. 1, OUP, Oxford.
Lewis, D. (1986), «Causal Explanation», en Id., *Philosophical Papers*, vol. 2, OUP, Oxford.
Loar, B. (1981), *Mind and Meaning*, Cambridge University Press, Cambridge.
Putnam, H. (1975a) «Brains and Behaviour», en Id., *Philosophical Papers*, vol. 2, Cambridge University Press, Cambridge, Mass. (Publicado originalmente en 1963.)
Putnam, H. (1975b), «The Nature of Mental States», en *Philosophical Papers*, vol. 2, Cambridge University Press, Cambridge, Mass. (Publicado originalmente en 1967.)
Ryle, G. (1949), *The Concept of Mind*, Hutchinson, London.
Searle, J. (1983), *Intentionality*, Cambridge University Press, Cambridge.
Searle, J. (1992), *The Rediscovery of the Mind*, MIT Press, Cambridge, Mass.
Wittgenstein, L. (1958), *Philosophical Investigations*, Basil Blackwell, Oxford. V.e.: UNAM-Crítica, México-Barcelona, 1988.

LA CONCEPCIÓN TELEOLÓGICA DE LOS ESTADOS MENTALES Y DE SU CONTENIDO

Daniel Quesada

I. INTRODUCCIÓN HISTÓRICA

A pesar de que en la reflexión filosófica acerca de las explicaciones en biología la utilización de un concepto de función con carácter teleológico tiene una larga historia, la relevancia de tal tipo de concepto para la filosofía de la mente y filosofía de la psicología sólo se comenzó a reconocer en la década de los 80.

Si es común al funcionalismo caracterizar a los estados mentales como *funciones* mentales, en general, el concepto de *función* que utilizan las teorías primeras y las más típicas teorías funcionalistas es disposicional-causal: el *papel funcional* que caracteriza a un estado mental, según tales teorías, coincide con sus efectos y causas, es decir, está constituido por el conjunto de estados y conductas que tal estado puede causar junto con el de estados y acaecimientos de los cuales es efecto, y la *función* del estado se caracteriza por los primeros.

Daniel Dennett fue uno de los primeros en señalar la importancia que, para la comprensión de lo mental, tiene concebir los estados y mecanismos mentales desde la perspectiva de sus «propósitos» o «fines» (cf. Dennett, 1971), pero concibe a éstos de un modo no-realista, relativamente a un «diseñador» que se los conferiría, llegando, al mismo tiempo que subraya la importancia que tiene contar aquí con la selección natural, a interpretar su papel meramente de manera metafórica, como si fuese el trabajo de un agente (la «madre naturaleza») que fuera la fuente última de tales «propósitos» (cf. Dennett, 1987).

Lo que en Dennett es sólo una postura instrumentalmente útil se convierte en William Lycan en una propuesta realista sobre la manera co-

rrecta de concebir los estados y procesos mentales. Se trata de identificar a los estados mentales haciendo referencia a su papel «en el fomento de los fines y estrategias de los sistemas en los cuales se dan» (cf. Lycan, 1981, 27). Lycan se adhiere al funcionalismo *homuncular* de Dennett, una concepción de las actividades mentales inspirada en el trabajo en Inteligencia Artificial, según la cual tales actividades han de estudiarse como si fuesen el resultado de la labor conjunta de múltiples «homúnculos» —múltiples mecanismos a diferentes niveles jerárquicos— cada uno dotado con su propia «misión» o «propósito», pero, a diferencia de Dennett, no concibe esta caracterización teleológica como el resultado de una mera manera de ver que presupone un intérprete o diseñador dotado de intencionalidad, sino que piensa que puede dársele un contenido realista y naturalista en términos evolucionistas (cf. Lycan, 1987, capítulos 4 y 5).

Lycan, no obstante, tendía a confundir la caracterización teleológica de un objeto con el paso a caracterizaciones más abstractas del mismo, haciendo de aquélla una cuestión de grado. Una llave, por ejemplo, puede caracterizarse —además de como llave— como una cierta colección de moléculas, un trozo de metal con un reborde dentado, un quitacerrajas, un abridor de puertas, algo que permite entrar en habitaciones de hotel, «algo que facilita las relaciones adúlteras [o como] un destructor de almas» (cf. Lycan, 1981, 32-33 y 1987, 43). Pero ninguna explicación en términos evolucionistas del concepto de función podría apoyar las últimas caracterizaciones. En la lista meramente se pasa de caracterizar el objeto en términos intrínsecos a hacerlo en términos relacionales, por sus efectos sobre un entorno concebido cada vez de modo más amplio.

Elliott Sober insistió con fuerza, en términos parcialmente similares a los de Lycan, en una reformulación teleológica del funcionalismo (cf. Sober, 1985), pero distingue de forma clara entre lo que es una función de un objeto o estructura —en el sentido en que el término se toma en biología— y lo que meramente es una manera más abstracta de caracterizarlo relacionalmente. No obstante, en ese escrito Sober se expresa escépticamente hacia las posibilidades del funcionalismo en último término, pues supone que el funcionalismo reformulado teleológicamente está abocado a la posición ingenuamente adaptacionista de suponer que cada estado o mecanismo de relevancia psicológica tiene una función a la que está bien adaptado.

La importancia de la variante teleológica del funcionalismo se ha hecho patente con las explicaciones funcionalistas-teleológicas de la intencionalidad del «significado» o «contenido» de los estados mentales—, de carácter completamente naturalista de Ruth Millikan y Fred Dretske. A ellas dedicaremos nuestra atención más adelante.

II. EL CONCEPTO TELEOLÓGICO DE FUNCIÓN

Como ya se ha mencionado, el contraste entre el funcionalismo teleológico y el no-teleológico se marca en torno al concepto de función. El segundo tipo de funcionalismo concibe las funciones como *disposiciones*, aunque no entendidas como disposiciones «puras», es decir, como meras regularidades (la propiedad disposicional que llamamos «solubilidad en agua» consiste tan sólo en que cuando ponemos un objeto soluble en agua se disuelve) al modo humeano, sino al modo realista que postula una causa estructural para tales diposiciones. El funcionalismo teleológico se basa, por el contrario, en un concepto «biológico» de función: no es lo que un estado o estructura hace de hecho, o lo que tiene, de hecho, capacidad para hacer (disposición) lo que determina su función, sino lo que se *supone* que ha de hacer.

Existen varias alternativas para dotar de contenido sin circularidad a esta idea (por lo demás totalmente general, no limitada a las funciones mentales), pero la más prometedora es la que apela a la *historia* del estado o estructura (para una opinión diferente cf. Bigelow y Pargetter, 1987). Podemos ver esta apelación implícita en la formulación de Larry Wright: «La función de *X* es aquella consecuencia particular de su estar donde está que explica por qué está ahí» (Wright, 1976, 78). Esto no es de ningún modo paradójico si entendemos por ello que cosas de tipo *X* han estado teniendo ciertos efectos causales que, por alguna razón (por ejemplo, porque son útiles o beneficiosos en algún sentido para los organismos en los que tales cosas se dan), han hecho que esas «cosas de tipo *X*» se «consoliden» o se sigan dando. Con lo que la apelación a la historia se hace explícita.

Podemos formular este concepto de función de la siguiente manera: *X* (cosas de tipo *X*) tienen la función *F* si y sólo si (1) *X* (cosas de tipo *X*) causa(n) *F* y (2) *X* (cosas de tipo *X*) existe(n) («están ahí») porque causa(n) *F*.

Como sugiere el paréntesis, la palabra «existe» debemos tomarla aquí en el sentido amplio apuntado en la formulación anterior que incluye la presencia consolidada de las cosas del tipo en cuestión, tal vez ínsitas en un peculiar entramado causal. Además, el tiempo presente empleado en la formulación es el presente intemporal, y la apelación a la historia, aunque de nuevo implícita, sigue haciéndose, pues no puede ser que la existencia presente de cosas del tipo *X* se explique por lo que esas cosas estén causando ahora. Se supone también que las afirmaciones de causalidad de la formulación anterior están implícitamente restringidas a ciertas condiciones normales que pueden determinarse sin circularidad: los *X* causan *F* cuando las circunstancias son parejas (*ceteris paribus*) a aquéllas que se daban en los momentos históricos en que el beneficio causado por su acción hizo que los *X* se «consolidaran».

Por lo demás, este análisis del concepto de función deja abierto el cómo se explica la existencia presente de cosas del tipo *X* por lo que esas cosas causaron en el pasado. Una posibilidad es que el proceso que llevó a la consolidación de los *X* en el pasado fuera un proceso de selección natural, siendo los *X* presentes reproducciones (probablemente en un sentido genético) de los *X* así seleccionados. Otra posibilidad adicional es que el proceso sea un proceso de aprendizaje y los *X* presentes sean nuevas ejemplificaciones de los estados a los que se llegó como resultado del mismo, los cuales, de este modo, vuelven a reproducirse (aunque el sentido preciso de la reproducción sería distinto al del caso anterior). Existen también otras posibilidades más complejas.

La formulación del concepto teleológico de función que se ha dado sólo es una aproximación de lo que podría ser una formulación más precisa. La versión más detallada y refinada de lo que parece esencialmente el mismo tipo de análisis del concepto es la de Millikan (1984). Millikan discute que el análisis de Wright tenga el componente histórico (1989a). En todo caso, es así como otros autores han interpretado a Wright y como lo hemos interpretado aquí.

El análisis anterior no implica que el lugar del pasado en que haya que buscar la «consolidación» o «selección» de los estados o estructuras que de ese modo adquieren una función sea el origen o la primera vez que se dio tal proceso de selección. La explicación de que un tipo de estados o estructuras siga reproduciéndose o dándose puede cambiar según el momento de tiempo que se considere. Puede que lo que explique hoy el mantenimiento de cosas del tipo *X*, aunque haya que buscarlo en el pasado, sea algo distinto que habría que buscar en un pasado más reciente a lo que hacía que esas cosas se mantuvieran o consolidaran en algún periodo anterior. Si ello fuera así, estaríamos ante un cambio de función. Los cambios de función son más frecuentes en estados o estructuras que resultan de procesos de aprendizaje, pero pueden darse también en estructuras o estados innatos que deben su existencia a procesos de selección natural (por ejemplo, se ha aducido que la función original de las plumas era la de contribuir a aislar el organismo del entorno, cambiando posteriormente a ser la de contribuir al vuelo).

El concepto de función así caracterizado posibilita distinguir entre lo que un estado o estructura de un cierto tipo hace efectivamente y lo que se supone que ha de hacer. Lo que se supone es aquello a lo que tales estados o estructuras deben su existencia (con vistas a las aplicaciones en filosofía de la mente, es importante observar que éste «se supone» es metafórico: la explicación dada del concepto de función permite prescindir de la relativización a un sujeto del «suponer»). Lo que tal vez efectivamente hagan o causen puede ser otra cosa (o nada en absoluto), y ello por diversas razones. El color de la piel de un camaleón que, tal vez como resultado de una enfermedad, tuviera los mecanismos de pigmen-

tación estropeados de modo que no coincidiera con el del entorno, seguiría teniendo la función (en el sentido analizado) de prevenir que el camaleón sea víctima de sus predadores, aunque, obviamente, no estuviera «funcionando» (en el sentido disposicional) de ese modo. En otras palabras, tenemos aquí una suerte de normatividad naturalista: hay estructuras o estados que «funcionan bien» (es decir, «actúan» de acuerdo con lo que es su función en el sentido analizado) y otros que «funcionan mal» (no «actúan» de acuerdo con lo que es su función).

El análisis permite también distinguir los resultados meramente accidentales del funcionamiento de un estado o estructura, de aquellos que constituyen propiamente su función, y ello incluso en los casos en que tales resultados accidentales sean beneficiosos para el organismo. Los sonidos que emite un corazón pueden contribuir al diagnóstico de una enfermedad del mismo, pero si esto no es lo que ha hecho que existan estructuras complejas como los corazones, emitir esos sonidos no es parte de la función del corazón (que puede describirse, a grandes rasgos, como la de bombear sangre).

Estas dos consecuencias del concepto teleológico de función le confieren ventajas sobre otros conceptos, en especial al aplicarse en filosofía de la mente.

III. VENTAJAS DE LA CONCEPCIÓN TELEOLÓGICA DE LOS ESTADOS MENTALES

Si, de acuerdo con el funcionalismo teleológico, reafirmamos que los estados mentales son estados funcionales, pero concebimos un estado funcional como un estado dotado de una función en el sentido «biológico» que se ha expuesto en el anterior epígrafe, ello nos proporciona varias ventajas desde un punto de vista teórico.

En primer lugar podemos dar cuenta directamente de la *normatividad* de los estados mentales. En *Investigaciones filosóficas*, Wittgenstein argumenta decisivamente contra la concepción mentalista clásica del significado (común a un gran número de pensadores, de Aristóteles al primer Wittgenstein, pasando por Locke y Saussure) señalando que es incompatible con el aspecto normativo del significado. Pues bien, aunque el funcionalismo de tipo disposicional esté alejado del mentalismo clásico, es sensible a una ampliación de la crítica de Wittgenstein, pues no puede dar cuenta del elemento crucial de normatividad que también está presente en los estados mentales. Si, por ejemplo, atribuimos a un animal la posesión del concepto de *verde* (o, al menos, un cierto concepto de *verde*) porque lo hemos adiestrado para discriminar el verde realizando acciones distintivas cuando está frente a algo verde, podemos decir que se equivoca en una ocasión determinada si entonces no realiza

(en circunstancias parejas) una acción de ese tipo frente a algo verde o realiza una acción de ese tipo frente a algo rojo. A nuestros mucho más complejos conceptos se les aplica también inmediatamente esa normatividad: *poseer un concepto* es ser capaz de aplicarlo al menos en casos claros o de rehusar su aplicación en casos en los que claramente no se aplica. Esencialmente lo mismo vale para otros estados mentales prototípicos como *tener una creencia*: hay creencias que son erróneas (falsas) y creencias que no lo son (verdaderas), y es muy diferente la contribución al fracaso o al éxito de las acciones en cuya realización intervienen causalmente esas creencias. El funcionalismo teleológico parece tener los recursos necesarios para dar cuenta de la normatividad de los estados mentales (más adelante veremos, a grandes rasgos, cómo esto se aplica a las creencias), pero no se ve cómo pueda hacerlo el funcionalismo disposicional.

El funcionalismo disposicional no requiere por sí mismo que haya un vínculo especial entre los estados mentales y el entorno. En cambio, los estados funcionales que resultan de aplicar un concepto teleológico de función son típicamente estados que un sistema no podría haber adquirido si estuviera aislado, sino que hacen referencia a rasgos del entorno (cuán proximales o distales sean éstos es una cuestión sobre la que volveremos en epígrafes siguientes). Ahora bien, una serie de autores (véanse Putnam, 1975; Burge, 1986; McDowell y Pettit, 1986 y Jackson y Pettit, 1988) han argumentado con fuerza en favor de la idea de que los estados mentales se caracterizan por su contenido «amplio», que incluye el entorno. Si ello es así, como parece (cf. Field, 1978; Fodor, 1980; Stich, 1983 y Devitt, 1989 para reservas sobre esa opinión), tendríamos aquí un nuevo motivo (aunque secundario comparado con el anterior) para preferir el enfoque teleológico.

Uno de los puntos fuertes del funcionalismo es que proporciona un marco abstracto y general en el que estudiar los sistemas cognitivos, puesto que un mismo sistema cognitivo puede estar concretado o realizado físicamente de maneras diversas. Esta característica la comparte el funcionalismo teleológico con el disposicional, pero el funcionalismo teleológico es capaz de suministrar un punto de vista aún más abstracto que puede resultar conveniente para el estudio de los sistemas cognitivos, toda vez que diversos mecanismos, cuya organización computacional o funcional-disposicional es también distinta, pueden tener la misma función (esto puede suceder parcialmente, por ejemplo, en los mecanismos para la visión de formas espaciales).

Se ha alegado además que el funcionalismo teleológico suministraría una razón adicional para el análisis funcional de un proceso cognitivo. Cuando lo que interesa es el estudio de un sistema concreto dado, parece que el punto de vista funcionalista perdería su razón de ser cuando ese punto de vista se concibe disposicionalmente, pero no cuando se concibe

teleológicamente. El punto de vista funcionalista así concebido suministraría aún una explicación del *porqué* del sistema, en el sentido de explicar qué es lo que hace que, a su vez, ese sistema exista. En tal sentido, el funcionalismo teleológico encajaría bien con la conocida propuesta metodológica de la distinción de niveles en ciencia cognitiva que se debe a David Marr (cf. Marr, 1982, cap. 1), y aun la reforzaría, suministrándole mayor apoyo teórico.

Como se ha dicho anteriormente, el funcionalismo teleológico puede suministrar un punto de vista más general y abstracto, pero ello no implica que las funciones que postula deban ser necesariamente generales y abstractas. Al contrario, parte del potencial del punto de vista teleológico deriva de la posibilidad que abre de ver un sistema dotado de una función global como compuesto por partes con funciones que propiamente han de desarrollar para contribuir a la función global y que han sido «seleccionadas» precisamente por esa contribución. De este modo, el funcionalismo teleológico puede recuperar la idea de funcionalismo «homuncular» de Dennett y Lycan, y la idea emparentada de análisis funcional de Cummins (cf. Cummins, 1975 y 1983), aunque no comparta los puntos de vista acerca de las funciones del primero y tercero de estos autores.

IV. FUNCIONALISMO TELEOLÓGICO Y CONTENIDO

Es en la explicación del contenido de los estados mentales donde el funcionalismo teleológico ha de mostrar su fuerza, y es en ese empeño en el que ha atraído la mayor atención. Característicamente, los estados mentales, como decía Brentano, se dirigen a algo o están dotados de «intencionalidad», es decir, tienen esencialmente un «significado» o «contenido». Dicho de otro modo, buena parte de los estados mentales son estados representacionales, son representaciones, y son las representaciones que, en último término, dotan a otras representaciones —expresiones lingüísticas, signos convencionales, códigos, prácticas culturales, obras filosóficas, obras literarias, obras pictóricas, esculturas— del contenido o significado que éstas puedan tener. La explicación del carácter representacional de los estados mentales constituye, por tanto, el reto fundamental para el filósofo de persuasión naturalista, pues la viabilidad de un enfoque naturalista de lo humano y de la cultura depende en definitiva de esa explicación.

Dentro de los intentos de «naturalización de la intencionalidad», el enfoque funcionalista sustituye al puramente informacional-causal. Según este último, la clave del contenido de los estados mentales se encuentra en el concepto de información: tales estados no son sino estructuras dotadas de contenido informacional, donde éste se explica por medio de relacio-

nes causales entre condiciones del entorno y tales estructuras (cf. Stampe, 1977 y Dretske, 1981).

El principal problema de este enfoque es que no da cuenta del elemento crucial de normatividad de los estados mentales al que aludíamos en el apartado anterior. Una dificultad donde el problema se manifiesta es la imposibilidad de descartar, desde ese punto de vista, contenidos disyuntivos indebidos (cf. Fodor, 1984). Por ejemplo, aun admitiendo que a la luz del día las representaciones mentales de gatos (percepción de un gato, creencias perceptuales sobre gatos) las causan invariablemente los gatos, en una noche oscura las pueden causar mofetas, por lo que, de acuerdo con el enfoque informacional-causal, parece que habría que admitir que el contenido de tales representaciones es la disyunción «gato-o-mofeta». En tal caso, uno no estaría cometiendo un error al tomar lo que percibe por un gato cuando lo que ha visto es una mofeta, pues sus aparentes percepciones o creencias sobre gatos son en realidad percepciones o creencias sobre gatos o mofetas. Con ello —generalizando el ejemplo— la posibilidad misma de que haya representaciones erróneas queda en entredicho. Pero si algo es claro acerca de las representaciones, es que la posibilidad de que una representación sea errónea forma parte del concepto mismo de representación. Con lo cual se hace patente que una teoría así no puede ser de ningún modo una teoría aceptable sobre las representaciones.

Aparentemente, una solución sería (volviendo al ejemplo) no contar las noches oscuras entre las ocasiones que son pertinentes para la determinación del contenido. Generalizando de nuevo, la solución consistiría en contar como pertinentes sólo las *condiciones normales*. El problema es entonces determinar sin circularidad cuáles son estas condiciones normales. Este es justo el problema que un enfoque causal-informacional no puede solucionar y para el cual, en principio, el enfoque teleológico parece especialmente preparado.

Varios autores han suministrado sus propias versiones del enfoque teleológico del contenido. Entre otros: Millikan, Stalnaker, Dretske, Fodor, Papineau y Matthen (cf. especialmente Millikan, 1984 y 1989b; Stalnaker, 1984; Fodor, 1990; Papineau, 1987 y Matthen, 1988). La mayoría de ellas incluyen elementos teleológicos en una explicación que básicamente no lo es (esto hace que pueda incluirse en la lista Stampe, 1977). Explicaré las ideas más relevantes con la atención puesta especialmente en tipos de estados mentales básicos y relativamente simples, como pueden ser las creencias perceptuales (creencias causadas por percepciones, como la creencia de que tengo un objeto redondo y rojo ante mí o que eso que acabo de ver es un gato).

Fodor (en la obra mencionada) supone que hay *mecanismos cognitivos* con la función de producir estados como las creencias, mecanismos que, en *condiciones óptimas*, funcionan como productores de creencias

específicas cuando y sólo cuando se dan circunstancias también específicas en el entorno. Es decir, producen (en tales condiciones) estructuras E que *covarían* con circunstancias C. Tales circunstancias constituyen el *contenido* de esos estados. Por ejemplo, en condiciones óptimas los mecanismos en cuestión producen en alguien la creencia de que ahí hay un gato (al modo de Fodor: ponen la representación «ahí hay un gato» —o, mejor dicho, su versión en el lenguaje mental— en el almacén de las creencias) si y sólo si (se da la circunstancia de que) hay ahí un gato.

El elemento teleológico en esta explicación es la apelación a condiciones óptimas, pues éstas son las condiciones en las que los mecanismos en cuestión actúan como se supone que deben actuar (o, si se quiere, como es su «misión» o «fin»). A su vez, lo que determina cómo se supone que deben actuar sería algún proceso natural, como un proceso evolutivo o un proceso de aprendizaje.

El propio Fodor (cf. Fodor, 1987, 104-106, que es posterior al trabajo mencionado anteriormente, aunque éste se publicara después) reveló un problema importante de su teoría teleológica del contenido. Según la teoría, se aducen condiciones óptimas para la actuación de los mecanismos productores de creencias, pero no parece que haya condiciones óptimas de aplicación general. Al contrario, las condiciones óptimas para observar un determinado tipo de objetos (objetos grandes, por ejemplo) son distintas de las condiciones óptimas para observar otro (objetos pequeños), por lo que se suscita la sospecha de que no es posible especificar tales condiciones óptimas si no es en relación con el tipo de objetos y propiedades sobre los que versan las creencias (los que constituyen su contenido), lo que haría circular la explicación.

En Dretske, encontramos también elementos teleológicos, esta vez superpuestos a una explicación de carácter informacional-causal. Algunas estructuras han sido «seleccionadas» (tal vez en el curso de un proceso de selección natural, tal vez en el de un proceso de aprendizaje) por el importante papel que juegan en recoger información para el organismo que —en el caso más claro— es vital o importante para la satisfacción de una necesidad biológica (cf. Dretske, 1986, 25). Cuando puede darse una explicación de ese tipo de la existencia de una estructura E, se dice que E tiene la función de llevar la información de que se trate, o la función de *indicar* la circunstancia en cuestión. Ésta constituye su contenido. De modo que el contenido C de una estructura E (lo que E representa) es lo que tiene la función de indicar, es decir, lo que efectivamente indica o la información que efectivamente tiene cuando el sistema en que se da E funciona normalmente y las demás condiciones son normales.

Quizás esta vez puede salvarse el escollo de la especificación sin circularidad de qué condiciones han de contar como normales, al no tener que darse de un modo general para un mecanismo productor de estructuras representacionales. Pero, como ha señalado Millikan, los estados y

estructuras naturales pueden, en condiciones perfectamente normales, indicar o contener información sobre muchas cosas; por consiguiente, según la explicación propuesta, éstas se deberían incluir entre lo que esos estados o estructuras representan o tienen la función de indicar, contra lo que patentemente es adecuado hacer. Por ejemplo, puesto que una cara roja puede, en condiciones normales, indicar o contener información sobre el hecho de que el que la tiene ha hecho ejercicio, o ha estado al sol durante un buen rato, o siente vergüenza, habría de decirse que el tener la cara roja representa o tiene como contenido todas esas cosas.

El problema de fondo en la explicación de Dretske (1986) es que no se ha explicado qué es lo que una estructura *hace* que pudiera constituir su función determinadora del contenido. La «función» de indicar o llevar información, en el sentido explicado, no es nada que la estructura haga o cause (y, por tanto, no se trata de una función en el sentido explicado en el apartado anterior). El problema queda oculto porque la palabra «indicar» tiene, junto a un sentido «pasivo», que es el que se recoge en la explicación anterior, un sentido «activo», que implica hacer algo. Pero este sentido activo queda sin explicar en el trabajo mencionado.

Algo parecido puede decirse de Matthen (1988). Allí se caracteriza la función de un estado de percepción como la de *detectar*: es un «estado que tiene la función de *detectar* la presencia de cosas de un determinado tipo» (1988, 20). Este tipo de cosas constituiría entonces el contenido del estado. Es probable al decir esto que se esté dando la pista de la buena dirección a seguir, pero no más que la pista, pues Matthen no explica en qué consiste la mencionada función. El problema es, por tanto: ¿qué es eso que hacen o causan las estructuras representacionales que constituye su función indicadora o detectadora?

Un paso adelante en la solución del problema, aunque se presente con rasgos inciertos, lo constituye Dretske (1988). Allí vincula Dretske la función indicadora con movimientos o conductas (en el sentido habitual del término; hay que advertir que una de las dificultades de la mencionada obra es la utilización de conceptualizaciones diferentes de las habituales, y ello a pesar de los esfuerzos que hace el autor por explicarlas).

De nuevo partimos de que están presentes en ciertos organismos estructuras E que ya son indicadoras de algo, en el sentido de que contienen información sobre alguna circunstancia o estado del entorno, digamos C. Supongamos que el organismo en cuestión se conduce de una manera específica M que le es «favorable» si se da C. Si ello sucede, puede ocurrir que en el organismo la estructura E llegue a estar asociada a la producción de ese tipo de conducta. Se «recluta» entonces a E para la producción de M, o «se fija» a E a la cadena causal productora de M.

El caso paradigmático de proceso en el que ocurriría lo que Dretske explica en su análisis es el del aprendizaje por condicionamiento operante (el modo en que una rata, por ejemplo, aprende a presionar un resorte

cuando se encuentra hambrienta, lo que tiene como efecto que consiga alimento). Al final de un proceso de este tipo (o de otros a los que pueda aplicarse el análisis), la estructura E es un factor causal en la producción de un tipo de conducta o movimiento M que es favorable, de algún modo, para el organismo en cuestión cuando se da la circunstancia C. Es decir, un tipo de conducta que es apropiado a la circunstancia C. Entonces es cuando puede decirse que E ha adquirido una función, justamente la función de producir (conjuntamente con algún otro estado interno) conductas apropiadas a C, debido, precisamente, a que contiene información acerca de C. Es entonces también cuando puede decirse que el contenido de E es C.

Dretske no explica exactamente así las cosas y sigue insistiendo de vez en cuando en que aquello que hace que una estructura tenga un contenido no está determinado por lo que esa estructura causa, sino por lo que la causa a ella (el sentido «pasivo» de indicar). Pero recurre crucialmente al movimiento o conducta causada por una estructura o estado interno para explicar cuándo adquiere un contenido y cuál es éste. De manera que parece razonable interpretarle como suministrando —aunque de un modo no enteramente claro— una explicación del sentido «activo» de la expresión «indicar».

Hay dos objeciones importantes a la teoría de Dretske. En primer lugar, no parece que la estructura indicadora tenga que existir previamente al proceso en que se la «recluta» para la producción de un movimiento, sino que puede perfectamente crearse en el proceso, quizás primero por azar, y luego manteniéndose por las consecuencias «favorables», tal vez por un proceso de aprendizaje. Este reparo y la ocasional insistencia de Dretske, mencionada antes, en que lo importante para la determinación del contenido es la causa de la estructura, parecen indicios claros de que Dretske está todavía demasiado inmerso en el enfoque informacional-causal. Pero la objeción de más peso es que no está claro cómo el análisis de Dretske puede resolver un problema obvio: los estados mentales, como las creencias, no están asociados a la producción de ningún tipo específico de conducta, sino que —en la acertada frase de Garrett Evans— están «al servicio de muchos proyectos». Dretske ha reconocido el problema y en la misma obra citada trata de indicar alguna vía general de solución, pero sus intentos no han resultado convincentes.

El carácter «activo» de lo que determina el contenido ha sido reconocido desde el comienzo por Millikan (cf. especialmente Millikan, 1984 y 1989b). Una forma de enfatizar ese carácter activo es señalar que una estructura o estado de creencia tiene como «consumidores» otros estados o mecanismos al cumplimiento de las funciones de los cuales tal estructura o estado *contribuye*, estando determinado su contenido por su contribución específica.

Más explícitamente, suponemos que los estados, estructuras o mecanismos «consumidores» de una estructura E de creencia tienen funciones en el sentido histórico causal explicado anteriormente. Si tienen funciones en tal sentido es que, al menos en algunas ocasiones, producen o han producido algo, algo que explica su existencia. Por ejemplo, los mecanismos productores del movimiento de la lengua de las ranas han provocado el movimiento de ésta con la consecuencia de que un cierto número de veces una mosca acaba siendo introducida en la boca de la rana. Concentremos ahora nuestra atención en tales ocasiones o situaciones relevantes. En ellas la ejemplificación de la estructura E coincide con cierta condición relevante del entorno sin la cual la acción de tales estados o mecanismos no realizaría su función. Pongamos que, en el caso de la rana, esa condición sea la presencia de una mosca situada de cierta manera frente a la rana. Pues bien, esta condición C del entorno que se da en las situaciones pertinentes, sin la cual *en esas situaciones* los estados o mecanismos que utilizan E (es decir, aquellos sobre los que E tiene un efecto causal, sus «consumidores») no realizarían sus funciones, constituye el contenido de E. Así pues, el contenido de una representación del tipo que estamos considerando («tipo creencia») es una condición normal (en el sentido histórico-causal explicado, no necesariamente en el sentido estadístico de ser frecuente) para la realización de las funciones de sus estados o mecanismos «consumidores».

Ahora bien, ¿cómo se distingue la condición normal relevante —la que constituye el contenido— de otras condiciones igualmente normales que deben igualmente darse para que los «consumidores» de una estructura representacional E realicen sus funciones? Sin duda la rana debe estar sobre un cierto soporte para poder proyectar la lengua de modo que atrape moscas; debe además haber oxígeno en el entorno; la temperatura ambiente debe estar entre unos ciertos límites; y un largo etcétera. Todas estas condiciones son condiciones normales en el sentido indicado. La condición normal relevante para la determinación del contenido se fija cuando atendemos a que E no es una estructura aislada, sino que forma parte de un sistema de representaciones cuyo contenido varía sistemáticamente con su forma. En el caso de la rana, por ejemplo, E corresponde a una cierta posición de la mosca, otra estructura representacional E' del mismo sistema corresponde a una posición diferente, y así sucesivamente. Lo que se llega a formar como resultado de un cierto proceso de selección natural o de aprendizaje es un mecanismo cuya función —de nuevo, en el sentido teleológico explicado— es producir representaciones, la correspondencia de las cuales con el entorno varía sistemáticamente con su forma, de un modo similar al cual las formas de las danzas de las abejas varían sistemáticamente con la ubicación del néctar (sólo que en este ejemplo, claro está, se trata de representaciones externas y los «consumidores» no son estados presentes en el mismo orga-

nismo, sino otros organismos del mismo grupo). Nada de esto se da con respecto al resto de las condiciones normales aludidas anteriormente.

Según la explicación anterior del contenido de un estado representacional E del tipo que estamos considerando, aquél no parece estar determinado por ninguna función de E, y, en efecto, Millikan ha afirmado en ocasiones que no es la función de E lo que determina el contenido: «Nótese que la propuesta no es que el contenido de la representación estribe en la función de la representación o del consumidor, en lo que éstos hacen» (Millikan, 1989b, 287; p. 89 en la reimpresión en Millikan, 1993). De este modo, la explicación del contenido de E, aunque profundamente inmersa en nociones del funcionalismo teleológico, no sería directamente funcional-teleológica. Sin embargo, parece que, al enfatizar lo anterior, lo que Millikan quiere descartar es únicamente que no se asocie el contenido con algo *específico* que las representaciones hagan (o hagan hacer) a sus consumidores. Una explicación de ese tipo sería, como hemos visto, muy limitada y, por tanto, problemática (cf. *ibid*, 289; p. 92 en la reimpresión).

En otro lugar, Millikan logra salvar la dificultad describiendo la función de una representación del siguiente modo: «... para representar la circunstancia C al consumidor (intérprete) I, la representación produce un cambio en I que adapta las ulteriores actividades de I a C, es decir, modifica las actividades de I de modo que las teleofunciones de I se lleven a cabo en, o *vía* la mediación de, o a pesar de que, C.» (Millikan, 1990, 156; pp. 128-129 en la reimpresión.) Por tanto, la función de una estructura representacional E (es decir, aquello que esa estructura hace que explica su existencia o consolidación) consiste en *adaptar* a sus consumidores a una cierta condición C. Por el modo en que Millikan describe este «adaptar», podemos decir que E contribuye (es un factor causal) a producir conductas que son apropiadas a una cierta condición C del entorno, condición que constituye su contenido. Además, Millikan vincula explícitamente su explicación de la función de un estado representacional E del tipo que estamos considerando al sentido «activo» del indicar que mencionábamos antes: «... hay otra cosa que podría ser indicar; podría ser algo que se le haga al consumidor de la representación por parte de la representación, a saber, el representar ciertas condiciones al consumidor (...) el provocar un cambio de cierto tipo en el consumidor» (*ibid.*).

Nótese, sin embargo, que en la explicación de Millikan no es en absoluto necesario que exista previamente en un organismo una estructura indicadora E que contenga información sobre una cierta condición del entorno C para que C llegue a ser (si se dan ulteriores circunstancias) el contenido representacional de E. Ni siquiera es necesario que la estructura representacional que llega a formarse como resultado de un proceso de selección o de aprendizaje sea una estructura indicadora, pues en tal

caso su ejemplificación debería corresponderse en todos los casos o en un porcentaje elevado de ellos con la presencia efectiva de la condición representada, y lo que encontramos usualmente es que los errores de representación son frecuentes. Incluso podría haber razones biológicas para que ello sea así, dado que un sistema representacional fiable tiene un alto costo y puede entonces ser más ventajoso poseer un sistema menos fiable pero con un costo menor (cf. Godfrey-Smith, 1991b).

V. CRÍTICAS Y ESTADO ACTUAL: CONCLUSIONES

El funcionalimo teleológico constituye una de las direcciones más importantes en que se desarrolla actualmente la investigación conceptual en filosofía de la mente. Sin embargo este enfoque debe hacer frente a dificultades considerables.

La primera dificultad relevante se mencionó de pasada en la introducción histórica, al hablar de la contribución de Sober. Otros filósofos además de Sober —notablemente Fodor (cf. Fodor, 1987, 105-106)— han afirmado que el funcionalismo teleológico se ve abocado a la posición adaptacionista insostenible de que todo estado o proceso mental innato o todos los componentes innatos de tales estados o procesos son beneficiosos para el organismo en que se dan. Una primera reacción ante esta dificultad puede ser señalar lo implausible que resulta el desarrollo de toda la estructura de creencias, deseos, intenciones, etc., sin suponer que esa estructura cognitiva tiene un «efecto estabilizante sobre el conjunto de genes» (Millikan, 1989b, 293-394; pp. 96-97 en la reimpresión; véase también Millikan, 1991, 155-158). Pero realmente no puede evitarse admitir lo inadecuado de una posición adaptacionista ingenua, y, por consiguiente, la necesidad de explicar la aparición de estados y procesos no adaptativos. El problema es que la permanencia de un rasgo que no es óptimo en lugar de uno óptimo no puede explicarse directamente por los efectos de ese rasgo. En último término, parece que el funcionalista teleológico se ve precisado a mostrar cómo su posición es consistente con tomar en consideración todas la fuerzas y factores evolutivos que operan para producir los estados y procesos mentales y cognitivos, y no sólo los optimizantes. Esto es algo que sólo recientemente se ha comenzado a hacer (cf. Godfrey-Smith, 1991a, capítulo 3), pero los inicios parecen prometedores.

La objeción más importante al funcionalismo teleológico es que la explicación del contenido que es posible hacer dentro de ese enfoque permite, después de todo, candidatos bien diferentes a ser los contenidos de una representación dada. Fodor, quien ha presentado esta objeción con particular fuerza, sostiene que, por ejemplo, cuando pensamos en las ranas y en sus intentos de capturar presas mediante el rápido movi-

miento de sus lenguas, una teoría teleológica nos deja sin poder decidir justificadamente si el contenido de las estructuras neuronales de las ranas cuando sus campos visuales resultan estimulados del modo pertinente son moscas o si son «cositas negras del entorno», si suponemos que, en las condiciones que se dieron en el proceso de selección del mecanismo las «cositas negras» en cuestión (o al menos un número suficiente de ellas) eran, de hecho, moscas.

La importancia de esta objeción resalta especialmente si se atiende al esencial vínculo que une el concepto de representación con el de representación errónea. En un caso como el anterior lo que está inmediatamente en juego es la posibilidad de determinación objetiva del error, sin introducir injustificadamente factores interpretativos del intérprete humano. Si Fodor está en lo cierto, no podría decidirse de un modo justificado si una rana que proyecta su lengua con un golpe seco cuando se le echan, en vez de moscas, perdigones, está cometiendo o no un error. En efecto, si afirmamos que el contenido de sus representaciones internas es «mosca», entonces la rana comete un error al proyectar su lengua, pero no lo comete si el contenido es «cosita negra (u oscura) del entorno». Si todo depende de cuál sea nuestra interpretación, tenemos que renunciar a la afirmación de que el contenido está objetivamente determinado para la rana misma, con lo que nos quedamos sin algo que tenga por sí mismo un contenido representacional, es decir, en definitiva, sin representaciones naturales.

Millikan ha afirmado que su teoría tiene una respuesta inmediata para el problema de la indeterminación del contenido, basada principalmente en su distinción del papel de los «consumidores» y «productores» de las representaciones. En el presente caso, puesto que lo que requieren los mecanismos «consumidores» de la representación para llevar a cabo sus funciones es que aquélla coincida con la presencia de moscas, o al menos de bichitos comestibles, ésta sería la condición normal para su adecuado funcionamiento, y, por tanto, el contenido de la activación de las neuronas relevantes de la rana sería «mosca» o, cuando menos, algo así como «bichito comestible» (cf. Millikan, 1991, 163), pero no simplemente «cosita negra del entorno». Sin embargo, la objeción de Fodor a esta réplica parece convincente: podríamos perfectamente decir que lo que los «consumidores» requieren son «cositas negras del entorno», dado que, en el medio ambiente en que se desarrollaron las estructuras neuronales en cuestión, un número suficiente de las «cositas negras» del entorno de la rana eran moscas o bichitos comestibles (cf. Fodor, 1991, 295).

Sin embargo, posiblemente la teoría de Millikan posee todos los elementos necesarios para una buena respuesta a la objeción de Fodor. En primer lugar, se habría de enfatizar que el proceso por el cual una representación se asocia con una condición del entorno es un proceso

causal, tanto si se trata de un episodio de selección natural como de uno de aprendizaje. Por consiguiente, y aun con la incertidumbre epistemológica propia de la complejidad de los casos implicados, estaría en principio ontológicamente determinado cuál es la condición que juega el papel causal relevante.

Esta línea de réplica debería ser aceptable para un realista intencional como Fodor. De acuerdo con ella, cuál sea el contenido de una representación natural en un caso concreto como el de la rana es una cuestión *empírica* muy compleja. Si una estructura representacional adapta a sus «consumidores» a una cierta condición del entorno, produciendo por consiguiente cambios en ellos que en una ocasión determinada conducen a cierto tipo de conducta, este tipo de conducta debería *variar conjuntamente con* la condición en cuestión (al menos cuando las circunstancias son como aquellas que se dieron cuando la estructura resultó seleccionada). Esta sería la base para excluir como candidatos a contenidos proximales, tales como los estados retinales, pues éstos no cumplen el requisito, ya que diferentes estados retinales pueden provocar, en ciertas circunstancias, la misma conducta y, a la inversa, el mismo estado retinal puede provocar conductas distintas (véase Millikan, 1984, 100). Igualmente tendríamos en las consideraciones anteriores el punto de partida para excluir contenidos «demasiado distales».

De este modo, desde el punto de vista teleológico al que nos referimos, deberíamos esperar que un cuidadoso análisis de los resultados de investigaciones avanzadas sobre sistemas cognitivos nos proporcionase afirmaciones bien determinadas sobre el contenido. Así, por ejemplo, podría seguramente argumentarse, atendiendo a las excelentes investigaciones de Konishi y sus colaboradores sobre el sistema perceptor de sonidos de la lechuza común (cf. Konishi, 1993, y, para más detalles, Konishi, 1986 y Konishi y otros, 1988), que lo que ciertos estados neuronales de esos animales representan es la presencia de sonidos de ciertas categorías y las direcciones de las que provienen, y no, por ejemplo, meramente diferencias en los tiempos de estimulación de las terminales nerviosas en uno y otro oído, ni tampoco, por otra parte, la presencia de tipos específicos de productores de tales sonidos (como las ratas).

Si las consideraciones anteriores no fueran suficientes, otro aspecto de la teoría de Millikan —sobre el que es imposible extenderse aquí— podría suministrar un elemento adicional importante (la sugerencia que sigue se debe a Manuel García-Carpintero, quien considera que el elemento en cuestión no sólo es importante, sino crucial para una teoría del contenido). En efecto, Millikan diferencia entre dos tipos de representaciones. Lo que caracteriza a las más complejas —dicho en los términos que hemos preferido para formular su teoría del contenido— es que adaptan a sus «consumidores» a los *mismos* objetos (cf. Millikan, 1984, 239-244). En otras palabras, el funcionamiento normal de los «consu-

midores» requiere una identificación de aquello que las representaciones representan, de su contenido. Millikan no enfatiza que identificar es siempre identificar como *el mismo X*, donde *X* es un tipo de objeto. Pero esta consideración conceptual puede suministrar un dato relevante que restrinja aún más el conjunto de candidatos a constituir el contenido de una representación (podría, por ejemplo, excluir «mosca» como contenido de las estructuras neuronales de las ranas en las situaciones pertinentes, pues no parece que podamos atribuir a una rana la capacidad de identificar moscas concretas, siendo que, para empezar, es incapaz de diferenciar las moscas de otros objetos bien diferentes).

Una condición necesaria para que se produzca identificación es que haya dos o más (usualmente serán muchas más) representaciones que se solapen y cuyo impacto en sus consumidores presuponga la «proyección» a un elemento común en lo representado. Pero este requisito muestra que, excepto quizás en los casos más sumamente simples de representaciones, tanto las representaciones como los contenidos representados son algo complejamente *articulado*. Ello nos recuerda que lo que usualmente encontraremos son sistemas de representaciones que requieren también sistemas de reglas de «proyección» o correspondencia. De esta suerte, una afirmación sobre el contenido de una representación tiene repercusiones sobre afirmaciones acerca del contenido de otras (hay pues un cierto recurso a totalidades, un cierto *holismo*, aunque éste no sea tan radical como en otras teorías del contenido; para la versión de Millikan cf., por ejemplo, Millikan, 1984, 107 ss.; de hecho, para Millikan no puede haber representaciones aisladas de ningún tipo; pero posiblemente sea éste un aspecto de su teoría que puede y debe modificarse). Es usualmente a través de una explicación unificada del éxito histórico de todo un sistema de «proyección» como hemos de justificar la preservación (mediante la apelación al correspondiente proceso de selección natural, o de aprendizaje, o quizás otros procesos) de las representaciones que constituyen un sistema.

Estas consideraciones podrían formar la base sobre la que construir una réplica a una serie de objeciones (relacionadas entre sí y también con la objeción de Fodor) a las teorías teleológicas del contenido, como la objeción de Block (cf. Block, 1986, 660) de que esas teorías «no explotan en absoluto la estructura composicional del lenguaje» (Block dirige esta objeción directamente contra la versión de Fodor de la teoría teleológica que se expuso en el apartado anterior; respecto a esta cuestión, la réplica, a partir de las mencionadas consideraciones, sería directa), o la objeción de que no se ve cómo pueden esas teorías dar cuenta del contenido de «representaciones teóricas» (cf. Block, *ibid.*, 658), o el problema del «contenido reducido» que Peacocke ha planteado, a saber —dicho brevemente— el de cómo podría justificarse desde el enfoque teleológico la atribución (correcta) de una creencia con un contenido determinado C en

lugar de la atribución (incorrecta) de una creencia cuyo contenido está constituido por todas las consecuencias de C que tienen un impacto causal en el agente (cf. Peacocke, 1990, 62).

Las dificultades que hemos considerado en este epígrafe son las principales responsables de un cierto ambiente derrotista de reciente cuño acerca del enfoque del funcionalismo teleológico. Sin embargo, como también hemos visto, no es en absoluto claro que el funcionalismo teleológico carezca de los recursos necesarios para hacerles frente. Por ello, ese enfoque funcionalista sigue siendo en la actualidad fuente de esperanza y expectativas para los que buscan una clarificación conceptual de la naturaleza de los estados y procesos mentales.

BIBLIOGRAFÍA

Bigelow J. y Pargetter, R. (1987), «Functions»: *Journal of Philosophy, 84*, 181-196.

Burge, T. (1986), «Individualism and psychology»: *Philosophical Review, 95*, 3-46.

Cummins (1983), *The nature of psychological explanation*, MIT Press, Cambridge, Massachusetts.

Dennett, D. (1971), «Intentional systems», reimpreso en Denett, 1978.

Dennett, D. (1978), *Brainstorms*, MIT Press, Cambridge, Mass.

Dennett, D. (1987), «Evolution, error and intentionality», en D. Dennett, *The Intentional Stance*, MIT Press, Cambridge, Mass.

Devitt, M. (1989), «A narrow representational theory of the mind», reimpreso en W. Lycan (ed.), *Mind and cognition*, Blackwell, Oxford, 1990.

Dretske, F. (1981), *Knowledge and the flow of information*, MIT Press, Cambridge (Massachusetts), 1981. V.e.: *Conocimiento e información*, Salvat, Barcelona.

Dretske, F. (1986), «Misrepresentation», en R. Bogdan (ed.), *Belief. Form, content, and function*, Clarendon Press, Oxford.

Dretske, F. (1988), *Explaining behavior. Reasons in a world of causes*, MIT Press, Cambridge, Mass.

Field, H. (1978), «Mental representation», reimpreso en N. Block (ed.), *Readings in philosophy of psychology, vol. 2*, Methuen, London, 1981.

Fodor, J. (1980), «Methodological solipsism considered as a research strategy in cognitive psychology», reimpreso en J. Fodor, *Representations*, MIT Press, Cambridge, Mass. 1981.

Fodor, J. (1984), «Semantics, Wisconsin Style»: *Synthese*, *59*, 231-250.

Fodor, J. (1987), *Psychosemantics. The problem of meaning in the philosophy of mind*, MIT Press, Cambridge, Mass.

Fodor, J. (1990a), «Psychosemantics, or where do truth conditions come from», en Lycan, W. (ed.), *Mind and cognition*, Blackwell, Oxford.

Fodor, J. (1990b), *A theory of content and other essays*, MIT Press, Cambridge, Mass.

Fodor, J. (1991), «Reply to Millikan», en B. Loewer y G. Rey (eds.), *Meaning in mind. Fodor and his critics*. Blackwell, Oxford.

Godfrey-Smith, P. (1991a), *Teleonomy and the philosophy of mind*, Tesis doctoral, Universidad de California en San Diego.
Godfrey-Smith, P. (1991b), «Signal, decision, action»: *Journal of Philosophy*, *88*, 709-722.
Jackson, F. y Pettit, P. (1988), «Functionalism and broad content»: *Mind*, *97*, 381-400.
Konishi, M. (1986), «Centraly synthesized maps of sensory space»: *Trends in neuroscience*, *9*, 163-168.
Konishi, M. y otros (1988) «Neurophysiological and anatomical substrates of sound localization in the owl», en G. Edelman, W. Gall y W. Cowan (eds.), *Auditory function, neurological basis of hearing*, J. Wiley & Sons, New York.
Konishi, M. (1993), «Audición binaural»: *Investigación y ciencia*, *201*, 26-33.
Lycan, W. (1981), «Form, function, and feel»: *Journal of Philosophy*, *78*, 24-50.
Lycan, W. (1987), *Consciousness*, MIT Press, Cambridge, Mass.
Matthen, M. (1988), «Biological functions and perceptual content»: *Journal of philosophy*, *85*, 5-27.
Millikan, R. (1984), *Language, thought and other biological categories*, MIT Press, Cambridge, Mass.
Millikan, R. (1989a), «In defence of proper functions»: *Philosophy of Science,* 56, 288-302. Reimpreso en Millikan, 1993.
Millikan, R. (1989b), «Biosemantics»: *Journal of Philosophy*, *86*, 281-297. Reimpreso en Millikan, 1993.
Millikan, R. (1990), «Compare and contrast Dretske, Fodor, and Millikan on teleosemantics»: *Philosophical Topics*, *18*, 151-161. Reimpreso en Millikan, 1993.
Millikan, R. (1991), «Speaking up for Darwin», en B. Loewer y G. Rey (eds.), *Meaning in mind. Fodor and his critics*, Blackwell, Oxford.
Millikan, R. (1993), *White queen psychology and other essays for Alice*, MIT Press, Cambridge, Mass.
Papineau, D. (1987), *Reality and representation*, Blackwell, Oxford.
Pettit, P. y McDowell, J. (eds.) (1986), *Subject, thought and context*, OUP, Oxford.
Putnam, H. (1975), «The meaning of "meaning"», en H. Putnam, *Mind, Language, and Reality*, Cambridge University Press, Cambridge.
Sober, E. (1985), «Panglossian functionalism and the philosophy of mind»: *Synthese*, *64*, 165-193.
Stalnaker, R. (1984), *Inquiry*, MIT Press, Cambridge, Mass.
Stampe, D. (1977), «Toward a causal theory of representation», en P. French, T. Uehing y H. Wettstein (eds.), *Midwest Studies in Philosophy,* vol. 2. University of Minnesota Press, Minneapolis.
Stich, S. (1983), *From folk psychology to cognitive science*, MIT Press, Cambridge, Mass.
Wright, L. (1976), *Teleological explanations*, University of California Press, Berkeley.

TEORÍAS DE LA ARQUITECTURA DE LO MENTAL

*Jesús Ezquerro**

I. INTRODUCCIÓN

Toda la empresa científica está basada en la idea, y quizá no siempre en la convicción, de que el mundo está formado por sistemas y mecanismos cuya conducta externa observamos. Desde este planteamiento, la ciencia suele operar postulando hipótesis acerca de los mecanismos internos que son responsables de la conducta externa a explicar [1]. La psicología no tendría por qué ser una excepción, y sin embargo ha presentado tradicionalmente algunas peculiaridades y escrúpulos filosóficos con respecto a su viabilidad como disciplica científica. Cuenta Daniel Dennett [2] que en el año 1978 se celebró en Tufts un debate, ya histórico, organizado por la *Society for Philosophy and Psychology*. En el curso de la reunión se entabló una fuerte disputa entre varios pesos pesados de la ciencia cognitiva, Noam Chomsky y Jerry Fodor, de un lado, y Terry Winograd [3] y

* Agradezco el apoyo del Proyecto de Investigación PS92-0041 del Ministerio de Educación y Ciencia (DGICYT), que asimismo me ha permitido pasar una fructífera estancia en el CSLI (Stanford University) donde este trabajo ha sido finalizado. También agradezco a Agustín Arrieta por haberse tomado el trabajo de leer el manuscrito y de hacer observaciones muy oportunas. Por último, agradezco a Fernando Broncano, coordinador de este volumen, y amigo, por la santa paciencia que ha tenido conmigo y los mimos que me ha dado.

1. Un problema diferente es el estatuto ontológico que uno está dispuesto a conceder a los mecanismos internos postulados, es decir, si se es realista o no, y de qué tipo. Pero este es un problema filosófico general que no impide que aun los más recalcitrantes enemigos de cualquier interpretación realista de las teorías científicas están de acuerdo acerca de que la ciencia debe proceder mediante la postulación de hipótesis acerca de los mecanismos internos de los sistemas cuya conducta se pretende explicar.
2. Dennett (1988).
3. Al parecer, T. Winograd no había llegado todavía a las posiciones que defiende en su 1986.

Roger Schank, del otro. Los primeros atacaban la inteligencia artificial frente a los segundos. Sostenían que una ciencia cognitiva desarrollada en los cauces de la IA convertiría a la psicología en algo carente de interés. A juicio de Chomsky, las dos únicas posibilidades interesantes serían, o bien hacer de la psicología algo parecido a la física, tratando de explicar sus regularidades en función de leyes profundas y matemáticamente elegantes, o bien admitir que los fenómenos psicológicos no están gobernados por leyes estrictas y que su estilo metodológico debería parecerse más bien al del novelista (psicología popular). Llegó un momento en el que Marvin Minsky, también presente en la reunión, terció en la disputa defendiendo que el camino correcto para esta ciencia no se encuentra en el dilema anterior, que, como ya se habrá adivinado, es una ejemplificación más de la tradicional polémica filosófica: «ciencias sociales o humanas» *versus* «ciencias naturales», sino que, para Minsky, la ciencia cognitiva era, sobre todo, cuestión de diseño, de «ingeniería». En contra de esta posibilidad se pueden encontrar argumentos parecidos también en J. Searle (1980), y no sólo en lo que concierne a su argumentación basada en el carácter puramente sintáctico de los procesos computacionales, o alternativamente, en su estupidez semántica, sino por su sorprendente conclusión de que la inteligencia es básicamente un asunto de materia, en este caso de materia neuronal.

Parece haber algo en el enfoque ingenieril de la mente que repugna a determinado tipo de humanistas, con independencia de sus inclinaciones materialistas o dualistas. Quizá por esa razón el debate en torno al estatuto metodológico de la ciencia cognitiva se ha venido planteando por los filósofos en términos del dilema anterior. Pero es muy probable que, como continuaba diciendo Dennett (*op. cit*), entre la mente como un cristal o una caja translúcida, y la mente como un caos, exista la posibilidad de considerarla como una «máquina», un objeto del que no cabe esperar que esté gobernado por profundas leyes matemáticas, sino un objeto de «diseño», analizable en términos funcionales. De este modo entra en escena la inteligencia artificial con el propósito de configurar el núcleo conceptual de la ciencia cognitiva. Pero obsérvese que lo hace de una manera un tanto oblicua (como una especie de tercero en disputa) con respecto a los planteamientos tradicionales en filosofía de la ciencia. No obstante, la cuestión acerca de si resultan posibles las arquitecturas cognitivas se ha hecho depender de la propia viabilidad de la psicología computacional, es decir, de la psicología que adopta la inteligencia artificial como núcleo conceptual y herramienta de trabajo.

Si, como hemos dicho, la empresa científica consiste básicamente en postular los mecanismos internos que dan cuenta de la conducta externa de los sistemas u organismos bajo estudio, éste *desideratum* se debería traducir, en el caso de la psicología, en el intento de dar cuenta, en general, de lo que nuestras mentes/cerebros hacen. No obstante, lo que

nuestras mentes/cerebros hacen (o creemos que hacen) se puede clasificar en dos grandes grupos de tareas: (*a*) aquellas que requieren manipulación consciente de información, como, por ejemplo, resolver problemas, hacer diagnosis médicas, jugar al ajedrez, probar teoremas en lógica y en matemáticas, etc., y (*b*) aquellas que hacemos todos los humanos adultos normales e incluso algunos animales también, con independencia de la base educacional, como por ejemplo, percibir el mundo del sentido común, hablar la lengua materna y entender las oraciones en dicha lengua que se emiten en nuestro entorno próximo, es decir, al alcance de nuestros oídos, etc. Es importante tener en cuenta que, típicamente, estas últimas habilidades no requieren control consciente, es más, con frecuencia escapan a cualquier tipo de control por nuestra parte, y tienen lugar de forma automática, o cuando menos, refleja. Así pues, desde esta perspectiva sumamente general y básica, el problema de la arquitectura cognitiva consiste en la búsqueda de un modelo de mecanismo que explique cómo resulta posible que sistemas u organismos como nosotros mismos sean capaces de exhibir una conducta tan variada.

Pero si descendemos a un plano un poco más concreto, comienzan a aparecer dificultades relacionadas con la disputa que se ha relatado anteriormente. Caben pocas dudas de que cualquier sistema capaz de exhibir la conducta mencionada tiene que ser, necesariamente, un sistema de procesamiento de información. Pero resulta que los ejemplos de sistemas de procesamiento de información que mejor conocemos son los sistemas computacionales. Luego es de esperar que, sea como fuere lo que con el tiempo lleguemos a averiguar acerca de cómo somos, si es que lo averiguamos alguna vez, casi con toda seguridad perteneceremos a la familia de los sistemas de procesamiento de información. Es decir, quizá se pueda poner en cuestión la pretensión de que cualquier sistema procesador de información es inteligente, pero parece no caber duda de que todo sistema inteligente es un sistema procesador de información. Por esa razón, o bien se admite que la inteligencia artificial nos puede ayudar a desentrañar la caja negra de los sistemas cognitivos, o bien es necesario renunciar a la posibilidad de desarrollar una ciencia de los organismos inteligentes. Chomsky y Fodor parecen preferir esta última opción, pero en el caso de Fodor viene acompañada por una argumentación en contra de las posibilidades de la inteligencia artificial para construir sistemas inteligentes, una argumentación que se enmarca, como veremos, en una propuesta de arquitectura cognitiva particular: la teoría modular de la mente.

Las cuestiones relacionadas con la arquitectura cognitiva son, en primer lugar, la propia caracterización de la noción, ya que ésta se usa en muchos sentidos diferentes que se pueden clasificar en dos grandes grupos: uno relacionado con la perspectiva desde la que se abordan los sistemas inteligentes, cuyo resultado son las arquitecturas parciales y las

descripciones a diferentes niveles; el otro tiene que ver con la noción de arquitectura general o unitaria. De estas últimas, o con pretensiones de serlo, hay muy pocas en oferta, y ello es un signo evidente de las dificultades de la empresa. En segundo lugar se plantea el problema de si son posibles o no las arquitecturas generales o unitarias (aquí es donde interviene el argumento de J. Fodor). Por último, aparece el díficil problema de cómo evaluar empíricamente las arquitecturas cognitivas. La importancia de la arquitectura para la ciencia cognitiva resulta obvia si se tiene en cuenta que, en principio, proporciona el marco dentro del cual tiene lugar todo el procesamiento de información, y por consiguiente es responsable de todas las restricciones fundamentales de la conducta inteligente del sistema, por lo que debería constituir el elemento central de cualquier teoría de la cognición. El gran problema, sin embargo, es que la arquitectura se halla oculta detrás del nivel de conocimiento, y sabemos que la primera explicación aparente a la mayoría de nuestras actividades inteligentes tiene que ver con deseos, objetivos, creencias y situaciones. La arquitectura cognitiva, sin embargo, aunque no sea suficiente para decirnos qué hacemos, sí que es indispensable para decirnos cómo y por qué conseguimos hacerlo.

II. LA NOCIÓN DE ARQUITECTURA

Antes hemos dicho que la arquitectura hace referencia a la estructura del mecanismo que hace posible la conducta inteligente de un sistema de procesamiento de información. Por consiguiente, la arquitectura corresponde al conjunto de las especificaciones fundamentales del sistema. Pero resulta que esta clase de sistemas se caracterizan por tener como ingredientes el par estructura-función: un sistema que tiene una estructura dada produce una conducta que depende de la realización de una función en dicha estructura. En este punto precisamente es donde entra en escena la noción de arquitectura. El término arquitectura se usa para indicar que la estructura de un sistema de procesamiento de información posee un carácter permanente u originario, con lo que habría que entender por arquitectura, en términos muy generales, la estructura fija que proporciona el marco dentro del cual tiene lugar el procesamiento cognitivo. Desde esta perspectiva, se podría afirmar que, en un sentido abstracto, los componentes de la arquitectura reflejan o representan las estructuras físicas subyacentes del sistema. Este es precisamente el significado técnico que la noción de arquitectura ha adquirido en las ciencias de la computación, donde arquitectura se refiere a la estructura del *hardware* que da lugar a un sistema que puede ser programado. Esto implica, cosa muy importante, que la arquitectura así caracterizada corresponde al diseño de un sistema que admite la distinción entre *hardware* y *software*. Esta

distinción ha venido siendo una tradicional fuente de confusión, tanto en lo que se refiere a la estrategia metodológica a seguir en ciencia cognitiva, como en lo que atañe a la cuestión acerca de qué puede incluirse bajo el término «arquitectura».

Como es bien conocido, la idea de que la ciencia cognitiva debe proceder tratando de caracterizar o hallar el «programa» (*software*) de la mente, haciendo caso omiso de la estructura que lo realiza, ha tenido importantes defensores, tanto desde el campo de la filosofía como desde la inteligencia artificial. La razón usualmente esgrimida ha consistido en el conocido argumento de la múltiple instanciación: si un mismo programa puede ser ejecutado en una variedad de estructuras físicas, entonces el conocimiento de dichas estructuras no nos puede aportar nada para la caracterización de los sistemas cognitivos [4]. Esta idea, sin embargo, está equivocada. La razón es que una teoría cognitiva desarrollada únicamente a nivel de conocimiento (similar al estilo del novelista, como decía Chomsky) no puede ser una teoría empírica en sentido estricto, puesto que resulta inmune, en la práctica, a la contrastación empírica. Podría ser, a lo sumo, una teoría psicológica de caja negra, pero para eso ya teníamos el conductismo, tan fuertemente criticado por el propio Chomsky. Ciertamente, la caracterización de las funciones cognitivas a nivel de conocimiento puede ayudar a entender cuestiones empíricas importantes, en la medida en que circunscribe y establece restricciones acerca de algunas propiedades de la clase de mecanismos capaces de ejecutar dichas funciones. Pero algo fundamental que hace la arquitectura es determinar la clase de algoritmos que son ejecutables *directamente* en ella, en lugar de ser sólo *simulables*, previa emulación de la arquitectura correspondiente, cosa, esta última, que también permite hacer cualquier sistema computacional de tipo general o universal, como son prácticamente todos los disponibles en el mercado. Esta distinción entre ejecutar directamente un algoritmo, o simplemente simularlo, resulta crucial para la determinación del carácter empírico de las teorías psicológicas y, por consiguiente, para el diseño de arquitecturas cognitivas. Pero también se encuentra en la raíz de la variedad de usos que se hacen de la noción de arquitectura.

El término «arquitectura» no es unívoco, y ello se debe principalmente a dos razones: una es de orden conceptual, y tiene que ver con los supuestos conceptuales de base a los que acabamos de aludir, acerca de cómo se debe proceder en la investigación en ciencia cognitiva; la otra está relacionada con las diferentes perspectivas desde las que se puede analizar un sistema cognitivo. Imaginemos que alguien desea comprar una casa para vivir, o bien encargar su construcción a un arquitecto y a una empresa constructora. Bien poca gente se conformaría con decir, por

4. Un clásico ya en este tipo de argumentación es Fodor (1974).

ejemplo, «quiero una casa con tres habitaciones, salón, cocina y baño». Usualmente queremos saber algo más, y no solamente porque puede haber una inmensa variedad de casas que respondan a esa descripción, sino también porque unas pueden ser muy distintas a otras en aspectos sumamente relevantes para nuestro interés. Por ejemplo, queremos saber cómo esté resuelto el problema de la comunicación entre las distintas piezas, y también el de la luz y la ventilación, qué tamaño tiene la casa en su totalidad y cada una de las piezas en particular, etc. Todas estas características suelen reflejarse en el plano. De modo que el plano nos da una descripción más aproximada de la casa que la primera descripción. Es decir, el plano nos ofrece mucha más información para decidir semejanzas y diferencias entre dos casas que la mera descripción funcional del número de piezas y para qué sirve cada una. No obstante, ni siquiera aquí acaban los datos que resultan de interés a la hora de comprar una casa. Existen otras características que los arquitectos suelen reflejar en el proyecto y que pueden hacer que dos casas sean muy distintas a pesar de que sus respectivos planos se podrían superponer de un modo completamente coincidente. Por ejemplo, queremos saber qué tipo de materiales se usan en las paredes y en el suelo de las diferentes piezas, si las paredes tienen materiales aislantes y las ventanas cámaras de aire, cuál es la estructura general del edificio, si es de tipo «jaula» o de muros de carga, etc. Como se podrá adivinar, todos estos aspectos son sumamente relevantes para decidir si dos casas son iguales, o si existe una buena correspondencia entre un modelo (o descripción) de una casa y una casa real. Por supuesto, alguien podría aducir que se puede conseguir una temperatura estable en la casa sin necesidad de introducir especiales propiedades aislantes en la arquitectura, por ejemplo, instalando un acondicionador de aire, o también podría decir que, a fin de cuentas, las dos casas con estructuras de sostenimiento diferentes se mantienen en pie igualmente en circunstancias normales. Pero se convendrá en que dos edificios que resuelven el problema de su propia estabilidad y el de la estabilidad de la temperatura de esas dos formas pueden ser muy diferentes entre sí. La razón es que su comportamiento puede ser muy distinto cuando se dan determinadas circunstancias, por ejemplo, cuando hay un corte de electricidad o un terremoto. Ignoro si el símil se puede prolongar mucho más, pero abusando quizá un poco de él, se podría afirmar que, a lo largo de la breve historia de la ciencia cognitiva, y particularmente dentro de la perspectiva que adopta la metáfora del ordenador, se han dado, y en cierta medida todavía se dan, todas estas posiciones a la hora de diseñar la estrategia metodológica y establecer criterios de correspondencia entre modelos cognitivos. Por otra parte, el símil puede dar igualmente una cierta idea acerca de cómo resultan posibles diversas descripciones de una casa dependiendo de la perspectiva o de los aspectos concretos que nos interese destacar, y lo mismo se aplica a la arquitectura cognitiva.

En realidad, el problema es algo más complicado en el caso de la arquitectura cognitiva que en el de la casa. Una de las formas de introducir las razones de la primera clase que ayudan a comprender la polisemia del término es teniendo presente la distinción, establecida por D. Marr (1982) entre los diferentes niveles a los que puede ser descrito un sistema cognitivo, aunque algunos prefieren hablar de niveles autónomos de organización, y este matiz no es inocente. Según Marr, los sistemas cognitivos son analizables a tres niveles diferentes: (1) el nivel computacional; (2) algorítmico, y (3) de implementación o de mecanismo físico. Marr se basó, para su teoría de los niveles, en el análisis de su propio trabajo y el de su equipo sobre el procesamiento visual, pero pronto se aplicó esta propuesta tripartita de niveles a los sistemas cognitivos en general. Desde este esquema se asume que los sistemas cognitivos admiten una descripción a tres niveles diferentes. El nivel más elevado sería el *nivel de conocimiento (o semántico)*. En este nivel se describe a los sistemas cognitivos como sistemas que poseen deseos, creencias, objetivos, etc. Dada esta descripción, se intenta explicar cómo estos sistemas realizan tareas inteligentes apelando a tales creencias y objetivos, *y asumiendo que están conectadas con sentido o incluso de forma racional*. Debe tenerse en cuenta, sin embargo, que esta caracterización del nivel de conocimiento es más afín a las tareas que al principio hemos dicho que requieren la manipulación consciente de información que a las que no. Para estas últimas encajaría mejor la denominación de nivel computacional que D. Marr emplea, dando a entender que a este nivel las tareas se describen en términos de función matemática o computacional. Para que los sistemas cognitivos puedan operar a ese nivel, es necesario que las representaciones a las que se apela estén realizadas o codificadas en un nivel a veces denominado *simbólico*, y otras, de modo más neutral, de *algoritmo*. En este nivel se funciona en términos de operaciones de procesamiento de información sobre esas codificaciones, lo que requiere que posean alguna estructura determinada. Este nivel a su vez tiene que ser realizado en términos de algún sustrato. Pues bien, lo que en el campo de las ciencias de la computación se denomina técnicamente «arquitectura», corresponde a este sustrato definido en algún lenguaje descriptivo adecuado. En los computadores ordinarios, este nivel corresponde al de registro-transferencia, donde vectores de bits son transportados de una unidad funcional a otra. En el caso de los humanos correspondería al nivel de circuitos neuronales, que admite una descripción abstracta como una red compleja, con un alto paralelismo, en donde se procesa un medio de señales continuas mediante conexiones excitatorias e inhibitorias. La cosa no termina aquí, pues el sistema admitiría un número indefinido de descripciones más básicas, como por ejemplo, a nivel de neuronas, moléculas, etc. Por otra parte, resulta que, incluso entre los tres niveles anteriores existe un número también indefinido de niveles que varían en concreción

y detalle [5]. Por esa razón, algunos autores prefieren hablar en abstracto de la relación implementación-especificación [6], de modo que esta relación se puede aplicar a cada par de una cascada de niveles: dados dos niveles cualesquiera, el más básico es siempre la implementación del menos básico, el cual se puede considerar como una especificación del primero.

Algunos autores, por ejemplo, Z. Pylyshyn (1984; 1989) afirman que esta disposición a tres niveles es lo que caracteriza a las arquitecturas clásicas frente a las conexionistas (una acepción, de uso muy frecuente, sobre todo en ámbitos filosóficos, del término «arquitectura»). Pero si se relativiza la noción de nivel en el sentido indicado anteriormente, la afirmación de Pylyshyn no puede ser correcta. De hecho, los sistemas conexionistas admiten igualmente la distinción a los tres niveles anteriores, junto con todos los niveles intermedios que pueda poseer cada sistema concreto, luego las diferencias arquitectónicas deben depender de otros aspectos [7]. No es el objetivo de este trabajo analizar las diferencias, virtudes y defectos de las arquitecturas clásicas y conexionistas [8], pero sí conviene poner de manifiesto un par de cosas. Primero, cuando se habla de arquitectura en este contexto, lo que está en juego no es el concepto de arquitectura general o unitaria, sino las propiedades que debe poseer la clase de algoritmos que implementa las especificaciones realizadas al nivel de conocimiento o nivel-1. Esta cuestión no estaba clara al principio de la disputa, hasta el punto de que para algunos, el núcleo central de la controversia residía más bien en la asunción por parte de los defensores de los modelos clásicos de la tesis de la autonomía del nivel de conocimiento con respecto a su implementación en niveles más básicos [9], cosa que explicaría la concesión que Fodor & Pylyshyn (1988) hacen en el sentido de que los modelos conexionistas podrían servir como «mera» implementación para las arquitecturas clásicas. Posteriormente, Zenon Pylyshyn (1989) ha reconocido abiertamente, como ya habían propuesto desde hacía bastante tiempo McClelland y Rumelhart [10], defensores

5. Esta clasificación de niveles atendiendo al doble parámetro de la concreción y el detalle, frente al único criterio de la concreción utilizado habitualmente, se debe a Carol Lynn Foster (1990), y tiene una importancia básica para probar la equivalencia fuerte entre sistemas.

6. Puede verse Swartout, y Balzer (1983), Mcclamrock (1991) y Galton (1993) para una defensa de este punto de vista acerca de la caracterización de los niveles y su relación.

7. Por ejemplo, T. Horgan y J. Tienson (1993) afirman también que existe una interpretación general de los tres niveles de Marr que resulta neutral con respecto a los supuestos fundacionales de los modelos clásicos, y por consiguiente, serían de aplicación a los modelos conexionistas, entre otras cosas porque, como mantienen estos autores, no existe tal cosa como «la concepción conexionista de la mente». El conexionismo, desde este punto de vista, consistiría más en un «modo de hacer las cosas en ciencia cognitiva» que en una concepción básicamente alternativa de la mente.

8. Puede verse, en este mismo volumen, la contribución de J. Corbí sobre el impacto filosófico del conexionismo.

9. Véase, para esta interpretación de la disputa entre modelos clásicos y conexionistas, Oaksford, Chater y Stenning (1990).

10. McClelland, & Rumelhart (1985) defendieron que el nivel de algoritmo está relacionado con

destacados de los modelos conexionistas, que la caracterización del nivel algorítmico constituye el objetivo central de la psicología cognitiva, ya que la mera equivalencia *input-output* no puede constituir, en modo alguno, la meta de la modelización en psicología. En segundo lugar, se habrá observado que la noción de arquitectura que se maneja en esta discusión no se refiere precisamente a la estructura del *hardware* que soporta el procesamiento cognitivo, sino que se refiere a algo mucho más intangible como es la caracterización del nivel de algoritmo.

Si a esta propiedad que tienen los sistemas cognitivos de poder ser descritos a diversos niveles, añadimos la enorme plasticidad que ofrecen los computadores convencionales, es decir, los diseñados con arquitectura von Neumann y similares, tendremos las pautas necesarias para concretar un poco más el terreno en el que se sitúa la noción, tan sumamente esquiva, de arquitectura. Esta plasticidad ha propiciado inconvenientes y ventajas. Entre los inconvenientes está el que, debido al hecho de que las arquitecturas computacionales disponibles son básicamente iguales, y a que son máquinas universales, es decir, que pueden ser programadas para computar cualquier función computable, ha existido una fuerte inclinación a hacer descansar el peso fundamental de la explicación de su conducta en el programa o *software*, ya que un mismo programa puede ser realizado por soportes (o *hardware*) muy diferentes en cuanto a su naturaleza física y a su estructura. Por esa razón se ha tendido a olvidar la importancia que tiene la implementación para la ciencia cognitiva, y a asociar la idea de computación y de algoritmo con la clase limitada de algoritmos que pueden ser ejecutados en este grupo limitado de arquitecturas. Pero este modo de proceder está equivocado puesto que, como ya hemos dicho antes, diferentes arquitecturas permiten ejecutar diferentes clases de algoritmos. Pensemos en el caso de la máquina original de Turing de codificación binaria. Por supuesto que nadie piensa seriamente que nuestras mentes/cerebros tienen una arquitectura similar a ésta, aunque en la práctica con frecuencia se olvida este hecho aparentemente obvio. Aunque se trate de una máquina universal, los expertos saben que sus procesos computacionales resultan extremadamente complejos, y que la complejidad de su secuencia de operaciones sufre variaciones dependiendo de cosas tales como la naturaleza del *input* y la tarea a realizar. Pues bien, estas variaciones son completamente diferentes a las que tienen lugar en las máquinas con arquitectura más convencional. Es un hecho bien conocido que en una máquina de Turing el número de pasos

aspectos centrales como la eficacia, la degradación natural (*graceful degradation*) y la actuación en condiciones de «ruido» (no ideales), entre otras, y también con la cuestión acerca de si un problema dado es difícil de resolver o no, qué problemas requieren más o menos tiempo para ser procesados, cómo se representa realmente la información, etc. Por consiguiente, parece evidente que, desde el punto de vista de la explicación, la implementación importa mucho para los conexionistas. Por ello precisamente sostienen que el nivel apropiado para hacer ciencia cognitiva es el nivel de algoritmo.

necesarios para recorrer una secuencia de símbolos se incrementa en proporción al cuadrado del número de secuencias almacenadas. Por el contrario, en las arquitecturas de registro o en las que tienen memoria de acceso aleatorio (en donde es una operación primitiva la recuperación de un símbolo por su nombre o por algún otro indicador), la complejidad del proceso de localización de un símbolo es independiente del número de secuencias almacenadas.

Antes nos hemos referido a la importancia de la distinción entre ejecutar directamente un algoritmo o simplemente simularlo; ahora se podrá comprender mejor el alcance de la distinción: si volvemos al ejemplo anterior, resulta que las máquinas de registro pueden ejecutar directamente determinados algoritmos que son imposibles en una máquina de Turing, a pesar de que estas máquinas son universales. Esto quiere decir que una máquina de Turing puede ser programada para ser «débilmente equivalente» a una de estas otras y computar la misma función que la implementada por el algoritmo en cuestión, pero para ello es necesario diseñarla para simular la secuencia de estados de la máquina de registro. Por consiguiente, la máquina de Turing estaría primero emulando la arquitectura de la máquina de registro, y por tanto ejecutando el algoritmo en la arquitectura simulada, algo completamente diferente a computar el algoritmo directamente. Como dice Pylyshyn (1989), esta distinción entre ejecutar directamente un algoritmo o hacerlo mediante emulación previa de otra arquitectura funcional diferente a la propia, resulta crucial para la ciencia cognitiva. La razón es que la distinción *apunta al problema central de qué aspectos de la computación pueden ser considerados literalmente como formando parte de un modelo (los supuestos ontológicos de la teoría) y cuáles pueden ser considerados como meros detalles de implementación* debido a nuestra necesidad de simular estos modelos en los computadores que tenemos a mano.

Por otra parte, como se podrá inferir de lo que acabamos de decir, la plasticidad constituye asimismo una de las ventajas más interesantes que el uso de la metáfora del ordenador ofrece para la ciencia cognitiva. Los ordenadores permiten simular sistemas cuyas operaciones tienen una naturaleza muy diferente de la de las máquinas en las que tales simulaciones tienen lugar, de modo que no hay necesidad de sentirse atados por la estructura concreta de los computadores convencionales, ya que éstos pueden ser utilizados con el fin de simular no solamente las actividades mentales de nuestros cerebros, sino también de simular dichas actividades simulando asimismo la estructura abstracta que creemos poseen nuestros cerebros [11]. Esta perspectiva supone invertir en cierto modo la metáfora del ordenador, es decir, usar los instrumentos de la computación para desarrollar la metáfora del cerebro. Por ello,

11. Ver Rumelhart (1989).

aunque los sistemas conexionistas deban ser simulados en máquinas con arquitecturas convencionales, ello no debe hacer perder de vista el alcance empírico de sus propuestas, aunque en este caso la hipótesis empírica acerca de la arquitectura mental/cerebral consista en una estructura abstracta que se emula en un computador convencional. Por ejemplo, la mayor parte de los modelos conexionistas se implementa en Lisp. Con el Lisp se define una arquitectura para un sistema procesador de listas: la memoria se organiza en estructuras asociativas, listas y listas de propiedades; los operadores básicos del lenguaje permiten crear y destruir listas, hallar un *item* en una lista, insertarlo, etc. Sin embargo, teniendo en cuenta la anterior distinción, esta circunstancia sería, como diría Pylyshyn, un mero detalle de implementación sin compromiso empírico alguno.

Hay que tener en cuenta, además, que el grado de detalle con que se especifique la arquitectura depende de la perspectiva que se adopte ante el sistema, es decir, del tipo de cuestiones que se pretendan abordar. Por ejemplo, desde una perspectiva básicamente neurológica se puede modelizar el cerebro como un sistema compuesto por neuronas, que a su vez son caracterizadas abstractamente como elementos binarios que pueden estar activados o desactivados y que pasan de un estado a otro con cierta velocidad. Pero si queremos responder a otro tipo de cuestiones, como, por ejemplo, la habilidad de razonar y responder a las situaciones del entorno, se puede modelizar el cerebro como un sistema formado por unos órganos sensores, una memoria a corto plazo, una memoria a largo plazo, y un sistema de emisión de respuestas. No obstante, como se adivina, una arquitectura de este tipo se puede especificar a su vez a diversos niveles de detalle y estructura de cada uno de sus componentes.

Esta disposición de los sistemas cognitivos en niveles y perspectivas es muy especial, como dicen Newell y col. (1989), es, después de todo, el ojo de la aguja a través del cual deben pasar todos los sistemas susceptibles de ser inteligentes. El problema es que hay una inmensa variedad de arquitecturas y sustratos físicos en los que estos sistemas pueden ser implementados, y no existe por el momento una idea clara acerca de esta doble variedad y sus consecuencias en la modelización y en la conducta. Por una parte, resulta muy difícil comparar arquitecturas desde la perspectiva de sus consecuencias en la conducta, debido a que, como hemos dicho, la arquitectura no es suficiente para determinar la conducta. Por otra parte, cada arquitectura suele asumir un marco general y básico propio, por razón de que siempre dan prioridad al tratamiento de determinados problemas. En el caso de las arquitecturas generales o unitarias, se da el problema añadido de que la contrastación empírica siempre tiene que ser, por necesidad, fragmentaria, en el sentido de que se someten a contrastación determinadas tareas concretas que sólo hacen uso de una

porción limitada de la arquitectura. Debido a estas complicaciones, el resultado final es un alto grado de inconmensurabilidad [12].

No obstante, la dificultad básica reside en el gran solapamiento existente entre las arquitecturas y los programas que ejecutan. Del mismo modo que el componente arquitectónico de una casa no determina cosas tales como, por ejemplo, su comportamiento en el mercado, ya que éste depende de factores externos tales como su situación, los planes urbanísticos municipales y la ley de la oferta y la demanda, la conducta de los sistemas cognitivos depende en un altísimo grado de las creencias, deseos, intenciones, etc., y el papel de la arquitectura se limitaría a posibilitar este comportamiento. Si hiciese esta labor de forma perfectamente ideal (en caso de que esta noción tenga algún sentido), la arquitectura constituiría un elemento neutro en la explicación de la conducta. Pero del mismo modo que en el mundo real no existen los planos sin fricción, ni espacio y tiempo absolutos, los sistemas inteligentes reales tienen fuentes de información limitadas, su propia memoria también lo es, deben tomar decisiones en tiempo real usando unos recursos de procesamiento limitados en cantidad y en velocidad, tienen *lapsus* lingüísticos y errores de memoria, etc. Muchos aspectos de la conducta pueden deberse incluso a factores más básicos, como el funcionamiento neuronal, la estructura molecular de los neurotransmisores, o hasta la temperatura y la ionización de la atmósfera. Todos estos efectos tienen que ver con la arquitectura y, desde esta perspectiva, la psicología cognitiva versa fundamentalmente acerca de la arquitectura.

Como afirman Newell y col. (1989), lo que proporciona la noción de arquitectura es el concepto de sistema global de mecanismos que se requieren para exhibir una conducta flexible e inteligente. Si esto es así, entonces una teoría de la arquitectura es una propuesta teórica para el mecanismo cognitivo en su totalidad. De ahí su importancia básica para la psicología cognitiva, y para hacer de ella una disciplina empírica en lugar de una disciplina puramente ingenieril o normativa. Desde este punto de vista, tanto la estructura y tamaño de una memoria, como su forma de procesamiento, los mecanismos de control implicados en la ejecución de tareas simultáneas (como son prácticamente todas), los mecanismos de recuerdo y ejecución de acciones, etc., suponen hipótesis empíricas acerca de la cognición. La arquitectura es, en definitiva, lo que hace del estudio de la conducta inteligente algo psicológico, en lugar de una mera reflexión acerca de la racionalidad de sus objetivos a la luz del conocimiento. Por ello, resulta imprescindible partir de lo que sabemos

12. Kintsch (1992) afirma, sin embargo, que el problema de la contrastación fragmentaria y sus consecuencias es algo intrínseco a las arquitecturas unitarias en la medida en que están diseñadas para cubrir un amplio rango de fenómenos. Así, lejos de ser un defecto, sería una virtud; después de todo, la gente hace un uso limitado de las capacidades cognitivas que tiene en cada tarea concreta.

acerca de las características de la conducta inteligente de los organismos (los humanos, por ejemplo) y de las restricciones que los diferencian de los sistemas capaces de computación universal.

III. REQUISITOS PARA UNA ARQUITECTURA COGNITIVA HUMANA

Una teoría de la arquitectura apropiada debería dar cuenta de las características, propiedades, modo de procesamiento y limitaciones de los sistemas cognitivos reales. He aquí unos cuantos de estos aspectos:

1. Comportarse de forma flexible en función del entorno.
2. Exhibir una conducta adaptativa (racional, orientada a metas).
3. Funcionar en tiempo real.
4. Operar en un entorno rico, complejo y detallado:
   a) percibir una inmensa cantidad de detalles cambiantes;
   b) utilizar grandes cantidades de conocimiento;
   c) controlar un sistema motor con muchos grados de libertad.
5. Utilizar símbolos y abstracciones.
6. Utilizar lenguajes, tanto naturales como artificiales.
7. Aprender del entorno y de la experiencia.
8. Adquirir capacidades en el curso del desarrollo.
9. Vivir de forma autónoma dentro de una comunidad social.
10. Exhibir auto-consciencia y sentido del yo [13].

La lista no necesita ser exhaustiva, pero en ella se muestran buena parte de los requisitos comprendidos, tanto en nuestra concepción ordinaria de los seres humanos como en la concepción informada por la ciencia. Parece indudable que el ser humano posee un sistema cognitivo flexible: ante la presencia de nuevos problemas ensaya posibles soluciones, ajusta las estrategias, reinterpreta, etc. Además posee una sofisticada capacidad de aprender que se manifiesta en casi todas sus actividades cognitivas: sabemos discriminar objetos complejos a partir de las impresiones visuales; discriminamos sonidos complejos a partir de las impresiones auditivas; aprendemos a comprender/producir lenguaje; a resolver problemas; a razonar, etc., y esta capacidad de aprender entra en juego cada vez que realizamos actividades cognitivas. También poseemos una gran capacidad generativa, ya que, nuestras mentes/cerebros no se limitan a repetir información, sino que, por el contrario, son capaces de generar una cantidad ilimitada de palabras, sentencias, expresiones aritméticas, etc. Somos capaces de hacer cada una de estas cosas y, lo más frecuente, hacemos varias a la vez, gesticulamos mientras hablamos, ha-

13. Puede verse desarrollada esta relación de capacidades en Newell *y col.* (1989).

blamos mientras conducimos, etc., es decir, el despliegue de nuestras actividades está caracterizado por un fuerte paralelismo.

Sin embargo, nuestro sistema de procesamiento de información está afectado de una peculiar serie de restricciones que una teoría adecuada de la arquitectura debería tener en cuenta:

1. En primer lugar, nuestra memoria a corto plazo tiene una capacidad muy pequeña si se la compara con las de los cumputadores que conocemos. El promedio de *items* nuevos que somos capaces de mantener en atención activa es aproximadamente de siete. A partir de este número, cada nuevo item que se añade empeora sensiblemente la capacidad de recordar y usar esa información. Hay que tener en cuenta, sin embargo, las características peculiares de estos *items* (habitualmente denominados *chunks*), puesto que se trata de fragmentos con sentido, de modo que una persona puede recordar más o menos con la misma facilidad siete letras al azar o siete palabras compuestas por varias letras, siete cifras al azar o siete números de teléfono (que tienen siete cifras cada uno). Debido a esta peculiaridad la gente puede aprender a recordar cantidades grandes de elementos inconexos a primera vista conectándolos mentalmente de forma que compongan fragmentos con sentido (mnemotécnica).

2. Una restricción especialmente importante es que el cerebro humano trabaja de modo muy lento si se lo compara con un computador. Éstos suelen ejecutar una operación básica cada nanosegundo (10^{-9} segundos), mientras que las neuronas operan en intervalos medidos en milisegundos (10^{-3} segundos). Es decir, la velocidad de los ordenadores viene a ser 10^6 más rápida. Si se supone (lo que probablemente sea demasiado suponer) que una neurona ejecuta una función más o menos equivalente a una instrucción elemental de computador, se sigue que el cerebro humano trabaja aproximadamente un millón de veces más lento que un computador. Las consecuencias de esta restricción son sumamente importantes para la arquitectura. La relativa lentitud de los circuitos neuronales implica que los procesos cognitivos humanos que se ejecutan en, más o menos, un segundo, suponen aproximadamente sólo cien pasos. Dado que la mayor parte de los procesos estudiados en percepción, recuerdo, comprensión y producción del habla, etc., consumen en torno a un segundo, parece razonable imponer como límite lo que Feldman (1985) denomina la «restricción del programa de 100 pasos» (100 *step program constraint*), lo que supone que debemos perseguir explicaciones de esos fenómenos cognitivos que no requieran más de cien operaciones secuenciales elementales aproximadamente.

3. Los humanos no somos buenos siguiendo reglas de modo preciso: tenemos lapsos de memoria, de atención, descuidos, falta de finura, de precisión, inexactitud. Quizá sea éste el precio a pagar por nuestra enorme flexibilidad, e incluso puede que esta aparente limitación, al ser con-

trolada, sea una de las fuentes de la creatividad, pero en todo caso, esta imprecisión contrasta fuertemente con el comportamiento de los computadores convencionales.

4. También a diferencia de los ordenadores, cuando los sistemas cognitivos humanos sufren fatiga, distracciones, accidentes o daños, no paran inmediatamente de funcionar. Siempre encuentran modos de continuar, por ejemplo, tomando más tiempo para resolver un problema, reduciendo la precisión o exactitud de las soluciones, etc. Esta peculiaridad es conocida en la jerga de la ciencia cognitiva como degradación natural o dulce (*Graceful degradation*).

Naturalmente, esta relación no es exhaustiva, y podríamos continuar tanto como la psicología y las ciencias del cerebro hayan podido averiguar o lleguen a averiguar en el futuro, pero es suficiente para hacernos una idea de las características y restricciones de los humanos como sistemas inteligentes. ¿En qué afecta todo esto al problema de la arquitectura? Como ya hemos dicho reiteradamente, la arquitectura por sí sola no determina la conducta, por lo que el problema que se plantea no es tanto el de la construcción de una teoría de la cognición que tenga modelos que respondan a estos requisitos. El problema estriba, más bien, en averiguar qué consecuencias tienen estos requisitos y restricciones con respecto a la forma que debe tener la arquitectura. Por ejemplo, parece claro que a la vista de que los procesos cognitivos son, de ordinario, extremadamente complejos, y teniendo en cuenta las limitaciones de tiempo señaladas, los algoritmos que se propongan deben operar con un alto grado de paralelismo. La razón es que cualquier computador serial, aun siendo capaz de simular una tarea, violaría la restricción del programa a 100 pasos incluso en el caso de los procesos más elementales. Aspectos como éste tienen, como se ve, repercusiones de hondo calado en el diseño de la arquitectura computacional humana.

Muchas de estas habilidades y limitaciones, a excepción, quizá, de las capacidades de vivir de forma autónoma dentro de una comunidad social y de exhibir autoconsciencia y sentido del yo [14], han recibido tratamientos particulares y propuestas parciales mediante teorías de la representación, reglas de aprendizaje, arquitecturas de procesamiento, etc. Pero cualquier actividad cognitiva real parece implicar la operación de muchas funciones simultáneamente. Cuando se procede solamente mediante la propuesta de modelos parciales y la contrastación psicológica se concentra en fragmentos particulares de conducta, cosa, en principio, difícil de evitar en la práctica, se corren serios riesgos que pueden producir un

14. En la actualidad, sin embargo, se está trabajando intensamente en los aspectos de interacción social dentro de la denominada inteligencia artificial distribuida, que pueden contribuir a la modelización del diálogo y otros aspectos de la interacción social. Con respecto al problema de la consciencia, puede verse, entre otros, D. Dennett (1991 y 1994).

conjunto inarticulado de teorías. Por ello es necesario instalar la investigación en el marco de arquitecturas globales, aunque ninguna tarea utilice todos sus aspectos. En todo caso, como se ve, la tarea es difícil.

A pesar de todos los velos que parecen cubrir la arquitectura cognitiva, existen algunas vías de aproximación. Probablemente, la más importante sea el tiempo. Newell (1990) ha puesto de manifiesto la importancia de las escalas de tiempo en los procesos cognitivos. En la escala de milisegundos es muy probable que estemos tratando con procesos automáticos que reflejan las propiedades de la arquitectura directamente. Un ejemplo útil al respecto son los experimentos de exploración de la conducta de respuesta inmediata a diferencia de la conducta más deliberada o controlada [15]. El problema es que, a medida que nos movemos hacia conductas más controladas, como por ejemplo razonar deliberadamente acerca de un texto que estamos leyendo, o resolver un problema matemático, van ganando importancia los elementos de conocimiento y estratégicos y nos vamos distanciando cada vez más de la arquitectura propiamente dicha. Lo ideal, como dice Kintsch (1992), sería proponer arquitecturas dentro de las cuales se puedan realizar ambos tipos de tareas. Más tarde veremos algunos ejemplos de propuestas de este tipo de arquitecturas generales. Otra estrategia para acercarnos a la arquitectura puede ser intentar observar las regularidades universales: cuando se observe alguna regularidad a lo largo de muchas variaciones en otros aspectos, también es probable que sea debido a la arquitectura. Una de las estrategias más viables en la práctica, sin embargo, es intentar construir arquitecturas experimentales que soporten los diversos requisitos. A pesar de los argumentos *a priori* que estamos acostumbrados a oír de parte de los filósofos, cada intento de diseñar una arquitectura, aunque finalmente fracase, suele ser positivo, porque ayuda a estudiar la naturaleza de los requisitos que se pretende que cumpla. Queda, finalmente, todo lo que nos pueda aportar el estudio de las estructuras neuronales creadas en el proceso evolutivo, un campo donde es muchísimo lo que podemos aprender todavía.

IV. ¿SON POSIBLES LAS ARQUITECTURAS COGNITIVAS?

1. *Argumentos quineanos*

Algunos autores discuten si realmente son posibles las arquitecturas. Por ejemplo, Anderson (1990; 1991) ha planteado la cuestión en el terreno de las posibilidades existentes para obtener teorías de la arquitectura con garantías de correspondencia en el mundo real, en este caso, con

15. Véase, por ejemplo, Schneider & Shiffrin (1977), y Shiffrin & Schneider (1977).

nuestro sistema cognitivo. Debido a la opacidad y al carácter velado de la arquitectura, como ya hemos anticipado, la empresa se torna bien díficil, más difícil, a todas luces, que lo que resulta obtener modelos teóricos en las disciplinas naturales que no tratan sus objetos como sistemas procesadores de información. Al objetivo general de tratar de hallar estructuras internas que expliquen la conducta, Anderson plantea dos tipos de obstáculos. El primero es el viejo y quizá intratable problema de la inducción, es decir, el problema de inferir la estructura de la caja negra de nuestras mentes a partir de la estructura de la conducta que produce. El segundo obstáculo es el problema, también aparentemente intratable, de decidir la arquitectura correcta, puesto que puede existir un número indefinido de propuestas acerca de la estructura mental compatible con las mismas consecuencias conductales. Anderson propone rebajar los objetivos de la ciencia cognitiva adoptando el enfoque racional. Este enfoque, afirma, puede ayudarnos a reducir el problema de la inducción en el sentido de que, si se asume que la conducta está optimizada con respecto a la estructura del entorno, y se conoce en qué consiste esta relación óptima, entonces obtendremos restricciones importantes sobre la clase de mecanismos que pueden implementar esta relación óptima. Por otra parte, Anderson también cree que el enfoque racional puede ayudarnos a reducir el problema de la decisión o identificación de la arquitectura correcta, puesto que proporciona una descripción a un nivel de abstracción previo a las propuestas de mecanismos. En este sentido, todos los mecanismos capaces de implementar la misma prescripción racional serán equivalentes a ese nivel de descripción. Llevando al límite este razonamiento, Anderson ha llegado a afirmar que no se necesita una teoría de la arquitectura [16], pero ya hemos visto a lo largo de este escrito unas cuantas razones de peso acerca de la necesidad de hacer teorías de la arquitectura.

De cualquier modo, este es un argumento general que, en principio, afecta a toda empresa científica en mayor o menor grado, pero no tiene por qué ser cualitativamente diferente en el caso de la ciencia cognitiva. Si para hacer ciencia hubiese que resolver previamente el problema de la inducción y el de la subdeterminación de las teorías por los datos, entonces no habría ciencia. Estos argumentos son muy familiares en filosofía, al menos desde que Quine los puso de manifiesto, pero sin por ello concluir que la ciencia es imposible. En adición a la tesis de la subdeterminación de las teorías por los datos, Quine ofreció el argumento de la indeterminación de la traducción, que añade restricciones adicionales para la viabilidad de las ciencias de la mente. Una posible lectura de este

16. Véase Anderson (1991a). Curiosamente, Anderson ha sido uno de los autores que más ha trabajado en el diseño de arquitecturas cognitivas. De hecho, es el responsable de una familia de ellas denominada ACT, que será descrita más adelante en este capítulo.

argumento podría ser que, mientras las ciencias naturales pueden funcionar sin resolver previamente el problema de la inducción, las ciencias de la mente no pueden. La razón es que mientras que las ciencias naturales no necesitan de científicos mecánicos para su desarrollo, las ciencias de la mente basadas en la concepción computacional necesitan inteligencia mecánica. Este es, en esencia, el argumento de Jerry Fodor que vamos a ver a continuación.

2. *La tesis de la modularidad (o el argumento en contra de la posibilidad de arquitecturas generales)*

Existen, por consiguiente, argumentos de orden cualitativo que ponen en cuestión la viabilidad de obtener teorías de la arquitectura que posean un carácter general o unitario, puesto que, según aducen, resulta imposible modelar los procesos centrales. El responsable de este argumento es, como acabamos de decir, Jerry Fodor. A primera vista puede parecer que la propuesta de arquitectura que hace J. Fodor (1983), con su teoría de la modularidad, es la de una arquitectura mental más, con todas sus peculiaridades. Sin embargo, como podremos comprobar, su argumento de fondo, y el más importante, es precisamente que la inteligencia artificial no tiene posibilidad alguna de modelar una arquitectura cognitiva general o unitaria.

En su aspecto positivo, la tesis de Fodor consiste en afirmar que las mentes están compuestas por facultades u órganos que tienen un carácter modular. Para comprender bien su propuesta, es conveniente recordar los dos grandes tipos de tareas inteligentes que hemos mencionado en la introducción. La tarea de elaborar criterios suficientes y necesarios para la definición de la noción de módulo parece algo imposible de realizar en la práctica. No obstante, Fodor (1983) propuso una larga serie de características que posteriormente han sido fuertemente criticadas [17] y, finalmente, en su (1987), decidió quedarse, al menos, con los siguientes cuatro criterios que él denomina «mayores»: *a) Especificidad de dominio*: Es decir, los módulos están especializados en el procesamiento de información de determinado tipo. Por ejemplo, es muy posible que poseamos un submódulo, dentro del módulo más general de la percepción del lenguaje, especializado en discriminar los «ruidos» (o sonidos) que corresponden a usos verbales de entre todos los ruidos en general que se producen en nuestro entorno. Este sería el módulo de análisis fonético. *b) Automaticidad*: los módulos ejecutan automáticamente sus funciones cuando son estimulados, sin que tengamos posibilidad alguna de control para evitar que procesen o computen. Por ejemplo, no podemos evitar entender una oración emitida en nuestra lengua materna si el hecho tiene

17. Ver Garfield (1987).

lugar en un entorno cercano dentro de nuestro umbral de audición, es decir, no la podemos percibir como un ruido sin sentido. *c) Encapsulación*: un sistema o módulo está encapsulado informacionalmente si solamente tiene acceso a la información representada dentro de las estructuras locales que lo sustentan. Por consiguiente, un sistema encapsulado (como son todos los sistemas computacionales) tiene acceso a una cantidad de información menor que la disponible en el entorno para el sistema u organismo al que pertenece, debiéndose esta limitación a motivos de su propia arquitectura. Por ejemplo, es muy plausible que nuestro sistema de percepción visual está encapsulado en este sentido. Pensemos en las ilusiones ópticas: no podemos evitar percibirlas de una forma determinada, aunque hayamos aprendido que son ilusiones (naturalmente, este aprendizaje tiene lugar por otros canales de información). *d) Rapidez*: sería una consecuencia directa de las tres propiedades anteriores y, en principio, esta propiedad parece ofrecer ventajas biológicas respecto a la adaptación al entorno.

Prácticamente todos los sistemas computacionales que conocemos comparten estas características. Resulta difícil saber cuántos módulos hay en nuestro sistema cognitivo, pero Fodor estima que existen fuertes razones para pensar que existen al menos dos grandes módulos (que a su vez pueden contener submódulos): el módulo de la percepción y el del lenguaje. Recuérdese que bastantes años atrás Noam Chomsky ya hablaba del órgano o facultad del lenguaje. En contradistinción a los sistemas modulares tenemos el sistema central. El sistema central es el sistema de la deliberación, la comprensión y la razón, el que recibe la información que le facilitan los módulos periféricos y la integra, caracterizándose precisamente por no poseer ninguna de las propiedades anteriores. Existen cuestiones abiertas acerca de las características que tiene la información que los módulos envían a los sistemas centrales, así como acerca del problema, muy difícil de resolver, de averiguar dónde termina el procesamiento modular (rápido, encapsulado...) y dónde comienza la deliberación. En cualquier caso, la propuesta de Fodor constituye, como se ve, un esbozo de arquitectura mental.

¿Qué consecuencias tiene este diseño arquitectónico? La primera y obvia es que, al menos en principio, debieran ser más fáciles de conseguir teorías acerca de la arquitectura de los sistemas modulares que de los centrales. Sin embargo, es un hecho bien conocido que los avances más importantes en IA han sido obtenidos en el terreno correspondiente al grupo de tareas cuyo procesamiento correspondería al sistema central. Concretamente, el desarrollo más exitoso realizado en IA ha tenido lugar en el campo de la planificación y la solución de problemas. Esta situación no tiene nada de sorprendente si tenemos en cuenta que, a primera vista, cuando se construyen sistemas de este tipo no es necesario cuidar de la plausibilidad psicológica de los modelos, porque la planifi-

ción y la solución de problemas son básicamente tareas normativas. En cambio, las restricciones impuestas por la adecuación empírica no se pueden dejar de lado cuando se construyen sistemas cuya conducta se basa más directamente en el trabajo de la arquitectura. La tesis de Fodor, no obstante, es que el empeño de la IA de modelar los procesos centrales está condenado al fracaso, ya que se tropezará con el problema del marco (*frame problem*).

Naturalmente, las ventajas que parecen ofrecer los sistemas modulares deben pagar un precio. Quizá el más importante sea el hecho de que sus propiedades convierten a los módulos en sistemas cuyo procesamiento es intrínsecamente irracional, puesto que al delimitar por imperativos de su propia arquitectura el acceso a información, queda reducido igualmente el espacio de posibles respuestas o soluciones que puedan ofrecer. Y aun éstas, siempre responderán a criterios arbitrarios determinados por la propia rigidez del sistema. El sistema central no parece tener esas desventajas, pero en cambio tienen otras como la lentitud. Al ser sistemas abiertos en su base de datos, tienen que deliberar antes de tomar decisiones y ejecutar acciones. Pero si emprendemos la tarea de considerar un conjunto no arbitrario de evidencia relevante disponible antes de decidir la fijación de una creencia o ejecutar una acción, nos toparemos de forma inmediata con el problema de decidir cuándo la evidencia considerada es suficiente, es decir, deberemos afrontar el problema de cuándo parar de pensar. Es el problema de Hamlet. Fodor piensa que el problema del marco es precisamente el problema de Hamlet contemplado desde la perspectiva del ingeniero: los expertos en inteligencia artificial tratan de construir sistemas racionales en el sentido de que sus mecanismos de fijación de creencias y de deliberación no están encapsulados. Pero al mismo tiempo pretenden que estos sistemas fijen alguna creencia o adopten una decisión de cuando en cuando, en lugar de quedar paralizados en su proceso después de haber intentado computar un conjunto infinito de evidencia disponible. El problema está en que, con objeto de evitar esta situación, se necesita algún modo de delimitar la búsqueda de evidencia. En los sistemas encapsulados, según hemos visto, la delimitación es arbitraria, pero por eso precisamente se dice que estos sistemas son irracionales por definición. Por consiguiente, el problema reside en cómo hallar una estrategia (un algoritmo) que delimite de forma no arbitraria la búsqueda de evidencia seleccionando únicamente la evidencia relevante. Fodor piensa que este problema no tiene solución.

Tradicionalmente, el problema del marco no se había presentado de este modo en los ámbitos de la IA, sino más bien como el problema de construir un robot sensible a las consecuencias de su conducta. Si un sistema actúa en el mundo real, sus acciones alterarán el mundo, como es obvio. Ahora bien, si el robot es racional, debería modificar sus creencias para dar cabida a las nuevas situaciones creadas como efectos de su ac-

tividad, sean efectos pretendidos o colaterales. Pero como sucede que esta revisión puede afectar a toda su base de datos y es necesario efectuarla después de cada acción, entonces el robot quedaría inmovilizado. La solución habitualmente aplicada a este problema ha sido el uso de una estrategia denominada «dormir al perro» (*sleeping-dog*). Esta estrategia consiste, brevemente, en acotar cada situación como una base de datos separada (referida a intervalos, en lugar de a instantes de tiempo) y dotar al sistema de algún mecanismo para inferir automáticamente la persistencia de lo que no cambia en cada acción que ejecuta, concentrándose únicamente en los cambios pretendidos por la acción. Fodor piensa, sin embargo, que esta forma de plantear el problema del marco y su solución lo subestima, porque o bien asume un supuesto metafísico fuerte y arbitrario, como es el de que cada acción o evento afecta de manera relevante a unas cosas sí y a otras no, o si se quiere reducir tal arbitrariedad se plantea el problema de la racionalidad y de la inferencia no demostrativa, es decir, nuevamente el problema de la inducción. Estamos, como se puede adivinar, en el punto que planteaba Anderson. Sin embargo, hemos visto que el problema de la inducción no puede paralizar la ciencia. Todo el mundo está de acuerdo en que el problema de la inducción no está resuelto, e incluso existen viejos y bien conocidos argumentos a lo largo de la historia de la filosofía en el sentido de que jamás se resolverán. Sin embargo son muchos menos los que están dispuestos a defender razonablemente que la ciencia es imposible, que no sirve para nada, o que no tiene sentido la idea de progreso. ¿Qué es lo que Fodor añade al argumento de la inducción, que pueda afectar en particular a la ciencia cognitiva? Sorprendentemente, como ya hemos anticipado, Fodor afirma que mientras las demás ciencias pueden progresar sin tener resuelto el problema de la inducción, la ciencia cognitiva no puede. Y no puede porque, mientras las demás ciencias no tienen como objetivo la construcción de agentes mecánicos, la IA tiene precisamente ese objetivo, por lo que sólamente puede aspirar a modelar los sistemas modulares, pero no los centrales, ya que éstos sufren del problema del marco [18].

Concedamos que el problema del marco, así planteado, no tiene solución. Ciertamente, incluso las soluciones parciales que se han pro-

18. Es importante tener en cuenta este razonamiento para entender por qué unas veces Noam Chomsky y Jerry Fodor aparecen como líderes indiscutibles de la concepción computacional de la mente y otras sus más acérrimos enemigos. La lectura de sus escritos se puede prestar a confusión en este sentido, pero se hace más comprensible si se ven a la luz del siguiente esquema conceptual: defienden la imagen computacional de la mente como un marco filosófico apropiado para entender las relaciones mente/cerebro a nivel ontológico (fisicalismo de casos), pero la rechazan en lo que concierne a la metodología de investigación que conlleva, es decir, el diseño de modelos computacionales, excepto para el caso de las facultades u órganos que, al estar encapsulados informacionalmente, tienen un acceso limitado a la información disponible, hecho que los convierte *ipso facto* en irracionales y, por consiguiente, inmunes al problema del marco.

puesto para abordar diversos aspectos del problema como el de la persistencia de los no-cambios, no sirven de mucho a la hora de evitar el problema del marco, o bien reducir sus efectos, en la modelización de la comunicación de agentes y la interacción social[19]. ¿Pero condena el problema del marco el intento de modelar la arquitectura mental a un mero ejercicio sin sentido? Hay dos modos de entender el argumento de Fodor, uno fuerte y otro débil. Según el sentido fuerte, el mero hecho de intentar averiguar la arquitectura mental carecería de sentido. Sin embargo, no se ve cómo este argumento no afecta a toda la ciencia en general. Si el problema de la inferencia no demostrativa es una barrera infranqueable a la hora de hacer ciencia cognitiva, ¿por qué no es un problema para toda la empresa científica en su conjunto? A fin de cuentas, todas las críticas a la propia noción de «progreso» científico, y al íntimamente conectado problema del realismo, tienen que ver con el problema de hallar un esquema racional que fundamente y determine la adquisición y el cambio de creencias. Pero este esquema es imposible de conseguir debido a parcialidad esencial de nuestra modelización del mundo real. Por esa razón casi todas estas críticas terminan afirmando que la solución al problema de la racionalidad y del realismo equivaldría a situarnos en el punto de vista del ojo de Dios[20]. De modo que el argumento de Fodor vendría a decir que para hacer ciencia en general no es necesario ser Dios, pero para hacer ciencia cognitiva en particular sí que es necesario ser Dios. Alternativamente, el argumento de Fodor parece implicar que, si tuviésemos una ciencia cognitiva acabada, entonces se terminarían los problemas de la ciencia, y la ciencia misma, ya que, conociendo a la perfección cómo conocemos, sólo necesitaríamos poner en marcha la maquinaria y el resto del conocimiento vendría por sí mismo, seguro, verdadero. Sin embargo, no parece que la ciencia haya seguido las rutas que divisó Laplace. Por otra parte, este argumento no dice nada acerca de si podemos o no practicar una ciencia cognitiva y una inteligencia artificial imperfectas, como todas las demás ciencias. Pero con este paso nos movemos a la interpretación débil del argumento.

19. Puede verse J. Ezquerro (en prensa) para una argumentación en el sentido de que las soluciones habituales aportadas al problema del marco no sirven de mucho en la modelización de planes multiagentes que requieren cooperación y comunicación. Desde la perspectiva de la persistencia y el cambio, es decir, desde el problema de crear un formalismo capaz de decir qué cambios produce una accción, sin tener por ello que enumerar explícitamente todas las cosas que no cambian, puede verse Martí-Oliet & Meseguer (1993). J. Shoham (1987) desglosa el problema del marco en dos subproblemas, el de la predicción (cuando se ha calculado una secuencia de operaciones para obtener un objetivo y éste no se da), y el problema de la cualificación (cuando se requiere una representación completa de todas las condiciones que tienen que ser satisfechas para la ejecución exitosa de una acción). Estas y otras muchas perspectivas del problema del marco han sido afrontadas usando diferentes técnicas. De todas formas, persiste una fuerte discusión conceptual acerca de la caracterización precisa del problema del marco. Como ejemplo, puede verse P. Hayes (1987).

20. Ver, por ejemplo, H. Putnam (1988).

Una forma de entender esta versión débil podría ser que, en un sentido básico, la indagación sobre la arquitectura cognitiva abarcaría los sistemas modulares y se pararía al llegar a los centrales. Pero, para que el argumento tuviese fuerza, Fodor debiera ofrecernos asimismo una clasificación clara de los procesos mentales, es decir, una línea netamente divisoria diciéndonos dónde termina el procesamiento modular y dónde comienza la deliberación racional. Ahora bien, mientras no contemos con criterios claros al respecto, tiene sentido seguir indagando acerca de la arquitectura mental, ya que no sabemos hasta dónde podemos llegar por esa vía. Al final, puede que los procesos estrictamente racionales sean bastantes menos de los que nos parecen.

La variedad de procesos cognitivos no parece corresponder a esta dicotomía neta, sino que hay buenas razones para pensar que existen niveles de procesamiento intermedio entre la mera transducción automática de las señales que nos llegan como estímulos y la deliberación racional. Pensemos, por ejemplo, en los procesos que se llevan a cabo en la comprensión del lenguaje en tiempo real, tanto en las comunicaciones orales como en la lectura de textos. Algunos autores como Kitsch (1992), cuyo modelo de arquitectura se expondrá brevemente más adelante, proponen que este tipo de procesamiento tiene que llevarse a cabo en una zona intermedia situada entre la percepción y la solución de problemas. Es muy improbable que el procesamiento sea puramente automático, porque ingredientes esenciales de la comprensión como el ligamiento variable para resolver referencias anafóricas, tienen una dependencia situacional y contextual, otras veces los textos no ofrecen suficientes recursos sintácticos para desambiguar, etc.; sin embargo, muy pocas veces nos paramos a realizar un proceso inferencial explícito en estos casos, lo que indica que el grueso del procesamiento tampoco puede ser solución de problemas. Los procesos de este tipo tienen lugar en la escala de milisegundos, cosa que los había convertido en intratables computacionalmente desde la perspectiva de la inferencia racional de la solución de problemas, porque también es intratable cualquier noción generalizada de inferencia. No obstante, Lokendra Shastri (1990; 1993) ha mostrado que este tipo de razonamiento implicado en la comprensión del lenguaje es tratable desde el punto de vista computacional. Lo que sucede es que no se había tenido en cuenta la íntima relación existente («simbiótica» la denomina este autor) entre la naturaleza de las representaciones que se manejan, la efectividad de la inferencia y la arquitectura computacional en la que tienen lugar los procesos. Ya hemos hecho alusión anteriormente a la importancia de la escala de tiempo como indicador de la cercanía o alejamiento a la arquitectura, o bien al conocimiento, como principales responsables de la conducta. Una opción para afrontar este problema es tratar de diseñar modelos específicos para cada tipo de procesamiento cognitivo. Sin embargo, una posibilidad más interesante es

tratar de hallar arquitecturas lo más generales y unitarias posible donde puedan caber muchos tipos de procesamiento, aunque no siempre utilicen todos sus componentes. En lo que resta de este capítulo expondremos algunos ejemplos de arquitecturas con esta pretensión de ser generales o unitarias. Aunque tendremos ocasión de comprobar que no son de hecho tan generales, tienen suficiente entidad como para mostrar su utilidad, *pace* Fodor, y con independencia de que sean correctas o no, que seguramente no lo serán.

V. EJEMPLOS DE ARQUITECTURA: SISTEMAS DE PRODUCCIÓN, ACT* Y CI

1. *Sistemas de producción*

1.1. Sistemas de producción y solución de problemas

La mayor parte de las arquitecturas cognitivas generales incorporan como un componente básico sistemas de producción, o simplemente producciones, bien en memoria separada, como es el caso de ACT*, bien integrados en una memoria única, como sucede en SOAR. Los sistemas de producción no pueden ser considerados, en sentido estricto, una arquitectura general, ya que fueron propuestos, en origen, con la intención de tratar la solución de problemas desde un punto de vista más realístico psicológicamente que los tratamientos habituales en inteligencia artificial. Por esa razón, conviene hacer una introducción al respecto.

Supongamos que una persona quiere ir al trabajo por la mañana y para ello utiliza su coche. Entra en el coche, acciona la llave de contacto y comprueba que el vehículo no arranca. Ante esta situación, si dicha persona está convencida que tiene una ignorancia absoluta de mecánica, lo normal es que se dirija al taller más cercano para que le solucionen el problema. Pero suponiendo que tiene algunas ideas acerca del funcionamiento de los coches, abrirá el capó del motor y comprobará, por ejemplo, si tiene las bujías en buen estado. Si tal es el caso, tratará de comprobar si entra combustible examinando la bomba. Si no observa nada anormal en la bomba, tratará de comprobar, por ejemplo, si alguno de los cables que suministran electricidad a las bujías está suelto. Si todos están bien conectados volverá nuevamente a comprobar si entra combustible al motor, examinando esta vez, por ejemplo, el estado del filtro, etc. Como se verá fácilmente, en cada uno de los pasos o comprobaciones que hace, el usuario del vehículo echa mano de conocimiento específico acerca del dominio, comprobando si se cumple determinada condición, para realizar la acción adecuada.

Esta forma de ver la solución de problemas encaja muy bien con una concepción de los humanos como sistemas de procesamiento de infor-

mación, y contrasta con los procedimientos de búsqueda en el espacio de estados para la solución de problemas. Estos últimos, que se muestran efectivos para solucionar cierta clase de problemas, no resultan los más apropiados para solucionar o describir otras clases de problemas. Por ello, un punto de vista alternativo es observar cómo la gente resuelve problemas y tratar de describir los estados mentales y los procesos cognitivos que tienen lugar en la gente cuando se encuentra realizando tareas de solución de problemas. Sin embargo, resulta obvio que es imposible observar los estados y procesos mentales de las personas directamente. La psicología y la ciencia cognitiva fueron durante mucho tiempo incapaces de tratar este fenómeno, quizá por la hegemonía del paradigma behaviorista hasta épocas muy recientes, pero sobre todo, debido a dos razones. La primera es que se carecía de técnicas sistemáticas para revelar los detalles de las representaciones internas de las personas; la segunda, que tampoco se tenía una explicación acerca de las relaciones entre el conocimiento y la acción, es decir, se carecía de recursos conceptuales para esclarecer cómo las acciones de los sujetos pueden ser controladas por las representaciones internas del mundo que poseen. Esta situación comienza a cambiar con las ideas de Newell, Shaw y Simon (1958; 1963) y, sobre todo, con el estudio seminal de Miller, Galanter y Pribram (1960). Estos autores introdujeron las nociones de «análisis de protocolos» y de «sistemas de producción», que proporcionaron las herramientas para analizar las representaciones internas de los sujetos y una explicación acerca de cómo las representaciones se influyen entre sí para producir la conducta.

Miller y sus colaboradores se inspiraron en el trabajo de los etólogos sobre la conducta instintiva de los animales: un animal «tantea» si se dan determinadas condiciones en el entorno, y si tal es el caso, entonces lleva a cabo la acción apropiada. Ahora bien, los humanos no se limitan a responder a las condiciones del entorno, también diseñan planes para guiar su conducta en la rutina de los problemas cotidianos. Un plan es un programa de conducta que se activa cuando se satisfacen determinadas condiciones. Supongamos que estoy planeando qué hacer la noche del sábado. Una posibilidad es ir a cenar a un restaurante, pero ello depende de que me ponga de acuerdo con unos amigos y de que éstos, a su vez, encuentren a alguien que se quede en casa al cuidado de los niños. Si estas condiciones no se cumplen, la posibilidad alternativa es ir al cine. Como se ve, se trata de un conjunto de planes interconectados, en el sentido de que dependen de que se cumplan ciertas condiciones para orientarse en un sentido o en otro. Resulta fácil de ver que estos planes pueden ser representados como reglas condición-acción, o más esquemáticamente, como árboles, o bien grafos *Y/O*. Un árbol o un grafo *Y/O* es el modo de representar soluciones para problemas descomponiéndolos en subconjuntos de problemas cada vez más pequeños.

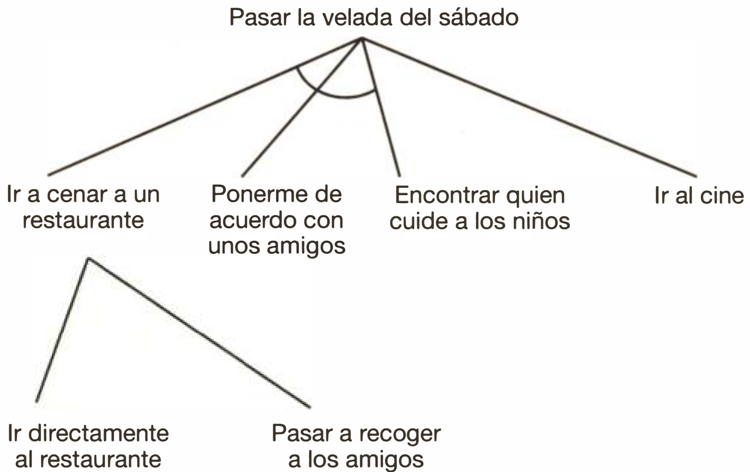


Figura 1.

Los arcos unidos por líneas curvas establecen vínculos Y entre los nodos, pues indican que para que se cumpla la condición principal, es decir, la estipulada por el nodo padre, debe cumplirse cada una de las sub-condiciones; por el contrario, los arcos no vinculados así representan una relación O entre nodos, indicando que el objetivo principal puede obtenerse intentando cualquiera de los sub-objetivos. La figura 1 representa un fragmento de un árbol Y/O que corresponde a mi plan para la velada del sábado. Su traducción al lenguaje ordinario vendría a ser: «Con el fin de pasar la velada del sábado puedo hacer O ir con unos amigos a cenar O ir al cine. Para ir con los amigos a cenar debo ponerme de acuerdo con ellos Y ellos tienen que encontrar quien cuide a los niños...».

Esta estrategia, dentro de la concepción de los agentes como sistemas de procesamiento de información, utiliza resultados psicológicos para hablar de estructuras cognitivas internas y de la estructura de la conducta humana, y no se limita a usar la estructura del problema solamente, como sucede con el enfoque clásico a la solución de problemas mediante la representación del espacio de estados del problema. Lo que se necesitaba para dar este paso es algún modo de «observar» los procesos del pensamiento humano. Como resulta evidente que estos procesos no están abiertos a la observación simple, hubo que recurrir a métodos indirectos, aunque, después de todo, empíricos. Uno de los métodos más profusamente utilizados es el de la medición de los tiempos de reacción. Otro se basa en la idea de que resulta posible obtener información acerca de los

procesos mentales de los sujetos basándonos en los informes de los propios sujetos acerca de lo que pasa por su mente mientras ejecutan una tarea. Este último método es conocido como análisis de protocolos[21].

El análisis de protocolos fue desarrollado por Newell y Simon como un método para estudiar los procesos mentales de los sujetos en la ejecución de tareas. Consiste en pedir a los sujetos que digan lo que pasa por sus mentes mientras resuelven problemas. Esta especie de «pensar en voz alta» se registra para su análisis pormenorizado hasta que el investigador consigue descomponer los protocolos de los sujetos en lo que se podrían considerar sus unidades de pensamiento atómicas, es decir, las operaciones mentales discretas que no pueden ser descompuestas, o bien no lo necesitan, para la comprensión del tratamiento que el sujeto hace de la tarea. Estas unidades atómicas son conocidas como «operadores». La aplicación de un operador cambia el estado de un problema de un estado de conocimiento a otro. Considerados en conjunto, los estados y los operadores aplicados a ellos constituyen el espacio del problema, y es dentro de este espacio, más bien que dentro de un espacio de estados dado previamente, donde tiene lugar la búsqueda de una solución a la tarea o problema. En resumen, un protocolo verbal adecuado permite al investigador diseñar una teoría abstracta acerca de cómo los sujetos resuelven problemas; la teoría se puede considerar entonces como un modelo de las actividades del solucionador de problemas y se puede traducir a un programa de ordenador.

1.2. La arquitectura de los sistemas de producción[22]

Acabamos de ver cómo los sujetos resuelven problemas examinando el estado actual del problema y escogiendo reglas apropiadas cuya aplicación transforma unos estados en otros en la búsqueda de una solución. Un sistema de producción es un medio para codificar este conocimiento basado en reglas; por ello, contiene tres componentes principales que corresponden a los estados de conocimiento, operaciones mentales y tomas de decisión del solucionador de problemas humano, respectivamente:

(1) *La memoria de trabajo*: es una base de datos cuyos contenidos representan lo que el sistema «conoce» acerca del problema en cada momento. Partiendo del estado inicial de conocimiento, éste se transforma

21. Necesariamente, este método debe ser complementado con otros. Los resultados experimentales han mostrado reiteradamente que la experiencia subjetiva es muy sospechosa como fuente de evidencia acerca de los procesos mentales. No obstante, resulta más fiable cuando se aplica a tareas como la solución de problemas que a otras tareas más básicas, puesto que muchas de estas últimas, por ejemplo, el *parsing* o el razonamiento reflejo, no proporcionan apenas consciencia subjetiva.

22. Puede hallarse una sencilla exposición en M. Sharples y col. (1989) así como en diversos manuales de introducción a la inteligencia artificial.

en un nuevo estado cada vez que un operador (regla de producción) le es aplicado. En los casos más simples, la memoria de trabajo se puede representar como una base de datos para hechos como:

[la jaula está abierta]
[Tweety está en la jaula]
[Garfield está en la habitación]
*
*
*

El tamaño de la memoria de trabajo es usualmente muy pequeño, del mismo modo que el número de *items* diferentes que los humanos manejan de ordinario en la memoria a corto plazo también lo es.

(2) *La base de reglas*: contiene el equivalente a lo que anteriormente hemos denominado «operadores», es decir, las operaciones «SI-ENTONCES» que modifican los estados de conocimiento. En los sistemas de producción se denominan «reglas de producción», o simplemente «producciones». Las condiciones, es decir, la parte «SI» se denominan a veces la «cabeza de la regla» (*rule head*), y las acciones, la parte «ENTONCES» se conoce como «el cuerpo de la regla» (*rule body*). Aunque aparentemente son como enunciados «si....entonces» de un lenguaje de programación, funcionan de modo diferente. Cada regla representa un *chunk* de conocimiento independiente que puede iniciarse, o activarse, cuando la condición entera coincide con items en la memoria de trabajo. Una vez activada la regla, sus acciones son ejecutadas, cosa que usualmente, aunque no siempre, supone modificar hechos, que dejan de ser verdaderos en la memoria de trabajo, y añadir nuevos hechos que se hacen verdaderos en dicha memoria. Una regla típica para la memoria de trabajo del ejemplo anterior podría describirse informalmente así:

Rn
Condición:
    (los contenidos anteriores)
Acción:
    modifica: [la jaula está abierta]
    añade: [la jaula está cerrada]

En lenguaje coloquial la regla anterior vendría a decir: «Si se da el caso de que la jaula de Tweety está abierta, mientras Garfield está en la habitación, y Tweety está dentro de la jaula, entonces cierra la jaula». Así pues, la activación de la regla cambia la base de datos de forma que los contenidos de la memoria de trabajo se transforman en:

[la jaula está cerrada]
[Tweety está en la jaula]
[Garfield está en la habitación]

En este ejemplo las acciones consisten simplemente en eliminar y añadir hechos en la memoria de trabajo, pero las acciones pueden producir muchas otras cosas, como veremos en la arquitectura ACT*.

Las reglas de producción no tienen orden, en el sentido de que resulta irrelevante el orden en el que se encuentran escritas físicamente en la base de reglas del sistema. La secuencia en la que son usadas depender del estado de la memoria de trabajo en cada momento. (Piénsese, por ejemplo, en una búsqueda en una diccionario donde la definición de cada palabra nos remite a otras, y éstas, a su vez, a otras, etc.) Si se activa más de una regla a la vez (es decir, si las cabezas de varias reglas coindicen con la base de datos), se usa una estrategia de resolución de conflicto para decidir qué regla debe ser activada, y existen varias estrategias de este tipo [23].

(3) *El interpretador del sistema de producción*: Es el programa que aplica las reglas. Ejecuta reiteradamente los pasos siguientes:

a) Satisfacción (*Match*): encuentra las reglas cuyas condiciones son satisfechas por los contenidos de la memoria de trabajo.

b) Resolución de conflictos: decide qué regla usar. Si no se satisface la parte condición de ninguna de las producciones, entonces el interpretador para.

c) Acción: ejecuta las acciones especificadas en el cuerpo de la regla.

d) Vuelve al paso (*a*).

El ciclo se repite hasta que, o bien no encuentra ninguna regla cuya parte-condición es verdadera en la base de datos, o activa alguna regla cuya parte-acción es *stop*. Si resulta afectada más de una regla, el interpretador debe decidir cuál de ellas se activa. La elección de estrategia de resolución de conflicto puede depender de la naturaleza de la tarea para la que está siendo utilizado el sistema en cuestión [24].

23. No obstante, hay algunos sistemas de producción en los que el orden puede ser importante, como sucede en el sistema PSG de Newell, utilizado para contrastar teorías psicológicas acerca de la memoria y el recuerdo humanos. En este sistema, las reglas se aplican en el orden en el que han sido escritas en la base de reglas, de modo que no se dan casos de conflicto entre ellas.

24. Posibles estrategias de resolución de conflictos son, por ejemplo, seleccionar la primera regla en la base de reglas que satisface los contenidos de la base de datos, o usar aquella regla cuyas condiciones para activarse incluyen las de todas las demás y algunas condiciones añadidas, bajo el supuesto de que es la más especializada; también se puede dar prioridad a reglas especializadas sobre las reglas más genéricamente aplicables, etc.

1.3. Características de los sistemas de producción

Los sistemas de producción tienen ventajas sobre los lenguajes de programación convencionales en el sentido de que resultan muy apropiados para su aplicación a tareas en las que el conocimiento puede cambiar o desarrollarse con el tiempo, y para aquellos problemas en los que el estado inicial y el final o solución pueden diferir de usuario a usuario. En general, los sistemas de producción se caracterizan por ser flexibles, modulares y psicológicamente plausibles.

La *flexibilidad* se muestra en que estos sistemas utilizan el mismo formato básico «SI-ENTONCES» para representar conocimiento en dominios muy diferentes. Ya se trate de inferir conclusiones a partir de ciertas premisas, o de ejecutar una acción dadas determinadas circunstancias, en todos los casos la parte-condición de la regla es contrastada con la memoria de trabajo y, en caso de ser satisfecha, se ejecuta la acción especificada en el cuerpo de la regla. La *modularidad* es una propiedad arquitectónica sumamente interesante que contrasta con la forma en que se comportan los sistemas ordinarios. En estos últimos, unos procedimientos llaman a otros, de forma que un cambio en un procedimiento puede implicar la modificación de todos los que lo llaman, y la simple eliminación de un procedimiento puede colapsar un programa completo. En contraste, las unidades funcionales de un sistema de producción (el conjunto de reglas de su base) constituyen *chunks* independientes de conocimiento, de modo que cada uno puede ser alterado o reemplazado sin incapacitar el sistema completo, y sin requerir siquiera la modificación de las otras reglas. Estas alteraciones pueden modificar o restringir el comportamiento del sistema, pero no lo detienen. La razón es que las reglas en un sistema de producción se encuentran separadas del programa que las ejecuta: es decir, no interactúan unas con otras directamente sino que lo hacen a través de los cambios en la memoria de trabajo. El flujo de control en la programación ordinaria se transmite secuencialmente de procedimiento en procedimiento, mientras que en los sistemas de producción vuelve a la base de datos de estado en estado. Debido a su flexibilidad y modularidad, los sistemas de producción han sido de mucha utilidad para el diseño de sistemas expertos, ya que estas propiedades los capacitan para ser ampliados y mejorados continuamente. Por otra parte, los sistemas expertos no necesitan almacenar la información únicamente en forma de reglas de producción. Hay sistemas que codifican su conocimiento en forma de red semántica; otros sistemas utilizan *frames*, o incluso alguna forma de la lógica de predicados. Finalmente, la *plausibilidad psicológica* se muestra, por ejemplo, en una capacidad tan «humana» como la de poder explicar por qué han adoptado una decisión o llegado a una conclusión. Los sistemas expertos que simulan la acción humana pueden ser preguntados puntualmente sobre las reglas que usan

e incluso acerca de las razones para usarlas. Los sistemas de producción pueden razonar, bien hacia adelante, de la evidencia inicial hacia la conclusión, bien hacia atrás, de una hipótesis hacia la evidencia que podría apoyarla, o bien en los dos sentidos. Los modos de razonamiento utilizados suelen venir determinados por el método usado por el experto humano[25].

Hay dos aspectos de la arquitectura de los sistemas de producción especialmente interesantes desde el punto de vista de su plausibilidad psicológica. El primero es la naturaleza dependiente de contexto de la satisfacción de producciones, a diferencia de las arquitecturas convencionales. Ello permite que un patrón pueda ser combinado con patrones contextuales adicionales formando un símbolo más complejo, lo que restringe la ocurrencia de la satisfacción cuando se encuentra presente el contexto[26]. El segundo aspecto concierne a la naturaleza del reconocimiento en la satisfacción de las producciones. A diferencia de las arquitecturas convencionales, en los sistemas de producción los símbolos son construidos a partir del mismo material que está siendo procesado para la tarea, de modo que el acceso a la memoria tiene una naturaleza asociativa, de reconocimiento o determinable por el contenido. Así pues, esta arquitectura responde a dos requisitos cognitivos importantes. Primero, la constante de tiempo de acceso al conjunto de la memoria obedece con bastante aproximación a los requisitos de tiempo real, propiedad que no tienen los sistemas de acceso secuencial como las máquinas de Turing. Las memorias de reconocimiento construyen los caminos de acceso a partir de los ingredientes de la tarea, y evitan, por consiguiente, actos deliberados de construcción, como se requieren en los esquemas de apuntadores de localización. Este puede ser realmente un requisito esencial de los sistemas de aprendizaje que deben desenvolverse por sí mismos completamente[27].

2. *La arquitectura ACT**

2.1. Motivaciones básicas: Modularidad *versus* sistemas unitarios

Según J. Anderson (1983), ACT* es una teoría de la arquitectura cognitiva, es decir, un modelo de los principios básicos de operación que ca-

25. Por otra parte, los sistemas expertos construidos como sistemas de producción pueden incorporar mecanismos de probabilidad, lógicas difusas o medidas de creencia para caracterizar las conclusiones. Puede verse, por ejemplo, el ya clásico manual de Charniak y McDermott (1985) para enfoques técnicos de razonamiento probabilístico y estadístico.

26. Pensemos en una situación en la que alguien está en la playa. Una producción podría ser «SI hace mucho calor, ENTONCES aplícate crema de protección solar». Ahora bien, una posible instancia de hacer mucho calor puede ser «hay más de 40 grados». Pero si se añade esta información, la acción a ejecutar podría ser «ponte a la sombra», en lugar de la anterior.

27. Ver Newell, Rosenbloom & Laird (1989).

racterizan a un sistema cognitivo[28]. De acuerdo con su denominación, el sistema está concebido para proporcionar una arquitectura unitaria (a diferencia de otras parciales o modulares) para el pensamiento o cognición de alto nivel. Es importante insistir en este supuesto de partida de que la cognición de alto nivel constituye un sistema unitario. Se trata de un supuesto que dista mucho de ser compartido en la comunidad cognitiva, como hemos visto, puesto que se opone a la conocida concepción modular de la mente defendida, entre otros, por Noam Chomsky y J. Fodor. Frente a esta concepción, Anderson mantiene que todas las funciones cognitivas de alto nivel pueden ser explicadas por un conjunto unitario de principios. En favor de esta tesis argumenta que, quizá a excepción del lenguaje, resulta absolutamente implausible sugerir que poseemos facultades u órganos especiales, producto de la evolución, para actividades tales como las matemáticas, el ajedrez, la pintura, la programación, etc. Los humanos llegamos a hacernos expertos en actividades de las que no es razonable pensar que han sido anticipadas en nuestra historia evolutiva. Sin embargo llegamos a adquirir estas habilidades gracias a una de las propiedades esenciales del intelecto humano, su plasticidad, que le permite adaptarse a tareas cognitivas muy diversas. Además, Anderson piensa que todas las actividades cognitivas de alto nivel tienen muchos aspectos en común, es decir, en ellas intervienen de forma difícilmente divisible las capacidades del lenguaje, solución de problemas, deducción, etc. Esta tesis unitaria es compatible, en principio, con la existencia de sistemas periféricos para visión, audición, etc., pero no con la tesis de que las funciones cognitivas superiores responden a diferentes órganos o facultades que se rigen por principios cognitivos particulares.

Si la interpretación que hemos hecho en la sección IV es correcta, existen razones para admitir una cierta compatibilidad entre la tesis unitaria de Anderson y la concepción modular, aunque posiblemente Chomsky y Fodor podrían no admitirlas en todos sus términos. Naturalmente, esta afirmación hay que entenderla en un sentido relativo, puesto que esta arquitecturas pretenden abordar capacidades como la solución de problemas, la acción racional o la comprensión del lenguaje. Este es el objetivo de las teorías como la de Anderson y, en general, el de todas (bien pocas, precisamente) las demás propuestas de arquitectura cognitiva existentes, como SOAR[29] y CI (*Construction-Integration*). Por

28. Anderson ha ido desarrollando y refinando a lo largo de más de veinte años toda una familia de esta clase de arquitecturas. El nombre responde a la expresión «*Adaptive Control of Thought*», y el asterisco (con el que se lee ACT «estrella») corresponde a una de sus versiones más evolucionadas y completas.

29. Ver Laird, Newell y Rosembloom (1987). SOAR no difiere sustancialmente de ACT*, excepto en lo que respecta a que cuenta con una sola memoria que incluye el conocimiento declarativo y

otra parte, cualquiera de los que proponen estas arquitecturas no tendría inconveniente en admitir, en principio, subsistemas o módulos para el lenguaje y la percepción en el sentido en que los propone Fodor. En realidad estos módulos serían los encargados de proporcionar al sistema la información que codifican a partir del mundo. Kintsch (1992) menciona expresamente la necesidad de complementar su sistema con un mecanismo de *parsing*, y la *codificación* en ACT*, como veremos más adelante, tendría que ser un mecanismo de este tipo, puesto que el sistema manipula información obtenida del exterior mediante algún tipo de codificación informacional. ¿Y qué otra cosa podría ser la facultad del lenguaje en el sentido de Fodor que un módulo de análisis fonético y un *parser*? El lenguaje y la percepción están presentes, en mayor o menor medida, en todos los procesos inteligentes de alto nivel, por lo que parece razonable pensar que, en caso de existir órganos o facultades, éstos cooperan de forma muy estrecha en la actuación inteligente, y ésta es una razón de peso en favor de un sistema unitario para dar cuenta de esta cooperación. No resulta plausible, por el contrario, proponer que la actividad lingüística tiene sus propias memorias de léxico y reglas, a largo y corto plazo, etc., y que con la capacidad de resolver problemas sucede otro tanto, y así sucesivamente. En todo caso, ACT* es un ejemplo de arquitectura en sentido global o propio del término, por lo que constituye en sí misma un compromiso empírico que, en principio, debe estar sujeto a contrastación experimental, lo cual es de suma importancia para el estatuto de la ciencia cognitiva como una ciencia empírica.

Por otra parte, uno de los aspectos cruciales de la cognición de alto nivel es el problema del control, es decir, el problema de qué es lo que da al pensamiento su dirección, y qué es lo que controla las transiciones de un pensamiento a otro. Como hemos visto, el uso en la arquitectura de sistemas de producción está relacionado con este problema. También sabemos que los sistemas de producción abordan el problema del control de un modo preciso y relativamente poco usual en psicología cognitiva. Ha sido frecuente encontrar otros tratamientos que producen modelos precisos para tareas específicas, pero que dejan sin tocar el problema de cómo el sistema procede de hecho en la ejecución de las tareas concretas. En los sistemas de producción, la elección de qué hacer en el paso siguiente tiene lugar mediante la selección de una regla de producción. Un problema básico en la elección, como también hemos visto, es el de las estrategias de resolución de conflictos. Con estas características, Anderson piensa que, finalmente, se ha dado con la clave para disolver el problema del homúnculo en psicología y filosofía de la mente.

procedimental, y en otros aspectos de menos importancia. Puede verse una comparación de las dos arquitecturas en Newell, Rosenbloom & Laird (1989).

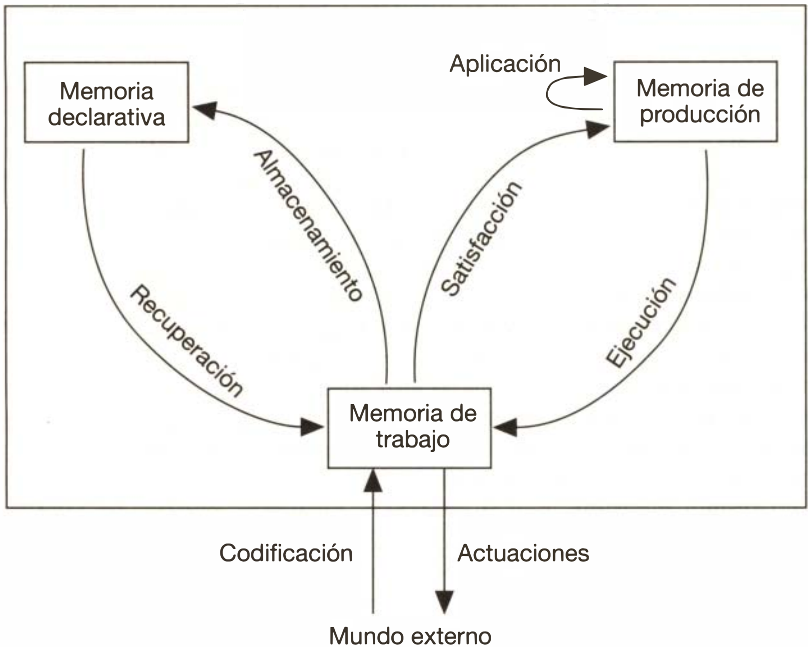


Figura 2.

2.2. El marco general ACT [30]

Toda la familia de modelos ACT se instala en el marco general que se puede observar en la figura 2. El modelo consta de tres memorias: una amplia memoria a largo plazo en forma de red semántica, una pequeña memoria de trabajo con *items* activos y un sistema de producción que opera sobre las memorias. La memoria de trabajo contiene la información que el sistema puede manejar a cada momento. Esta información es tomada de la memoria declarativa y se complementa con las estructuras creadas por los procesos de codificación de información que provienen del exterior y con las acciones que llegan de la memoria de producción. Como se ve, la memoria de trabajo se encuentra implicada prácticamente en todos los procesos del sistema. Los procesos de codificación depositan información sobre el mundo exterior en la memoria de trabajo, y los procesos de actuación convierten comandos de la memoria de traba-

30. Para la presentación del modelo seguir aquí la propia descripción que hace Anderson (1983), tratando de ahorrar al lector detalles técnicos, como las ecuaciones que rigen el establecimiento de la fuerza asociativa entre nodos y la difusión de activación.

jo en conducta. Estos dos procesos constituyen la periferia del sistema, o su contacto con el mundo exterior.

Los procesos de *almacenamiento* crean grabaciones permanentes de los contenidos de la memoria de trabajo en la memoria declarativa. Los procesos de recuperación retoman información de la memoria declarativa para pasarlos a la memoria de trabajo. Los procesos de *satisfacción* ponen los datos de la memoria de trabajo en correspondencia con las condiciones o cabezas de las reglas de producción. Los procesos de *ejecución* depositan las acciones de las reglas de producción cuya condición es satisfecha en la memoria de trabajo. Finalmente, los procesos de *aplicación* se reciclan en la memoria de producción, reflejando el hecho de que han sido aprendidas nuevas producciones a partir del estudio de la historia de aplicaciones de las reglas de producción existentes. Por ello, hay un sentido básico según el cual, los modelos ACT son teorías procedimentales del aprendizaje, es decir, aprenden actuando.

Sin embargo, un marco general de este tipo, aunque reúne los bloques básicos de la estructura arquitectónica de un sistema cognitivo, no llega a constituir una teoría, debido a que no propone, en sentido estricto, hipótesis contrastables que puedan ser sometidas a la prueba de la explicación y la predicción empíricas. Para situarse a este nivel, la teoría debe especificar, según Anderson, los siguientes puntos:

1. Las propiedades representacionales y las consecuencias funcionales de las estructuras de conocimiento de la memoria de trabajo.
2. La naturaleza de los procesos de razonamiento.
3. La naturaleza de los procesos de recuperación.
4. La naturaleza de la aplicación de producciones, que a su vez se subdivide en:
   a) el mecanismo de satisfacción de patrones (*pattern matching*);
   b) los procesos que depositan los resultados de las acciones de producción en la memoria de trabajo;
   c) los mecanismos de aprendizaje mediante los que la aplicación de producciones afecta a la memoria de producción.

2.3. Supuestos de ACT*

1. La hipótesis básica de ACT*, desde el punto de vista arquitectónico, es *la distinción entre conocimiento declarativo y procedimental*. Anderson mantiene que esta distinción proporciona importantes ventajas al sistema porque permite separar la memoria declarativa de la procedimental. Esta separación facilita la salida de las situaciones de resolución de conflictos, puesto que la recuperación de datos desde la memoria declarativa no tiene que competir con las producciones que ejecutan la tarea. Los

sistemas de producción clásicos utilizan reglas de producción para recuperar información declarativa, y estas reglas tienen que competir con las que ejecutan la tarea, y el problema de la competición se agrava debido a que la resolución de conflictos en los sistemas clásicos ha tendido a permitir únicamente la aplicación de una sola regla de producción. Este problema no sería excesivamente grave si la cantidad de información relevante para la tarea fuese pequeña, pero los resultados experimentales en *priming* asociativo (el efecto de la presentación de una información sobre el acceso a información asociada) muestran que la cantidad de información traspasada a la memoria de trabajo es muy grande. Además, la separación entre estas dos memorias permite reducir, según Anderson, la gran diferencia en cuanto al tiempo consumido para el almacenamiento de información declarativa y procedimental que tienen los sistemas de producción clásicos.

2. *Representación declarativa*: el conocimiento llega en *chunks* o *unidades cognitivas* (esta es la denominación que reciben en ACT*). Las unidades cognitivas pueden ser cosas tales como proposiciones (como [ama, Isabel, Pedro]), filas [uno, dos, tres], o imágenes espaciales (un triángulo dentro de un círculo). En todos los casos una unidad cognitiva codifica un conjunto de elementos que se encuentran en una relación determinada. Los *chunks* no contienen más de cinco elementos, aunque se pueden crear estructuras más complejas mediante estructuras jerárquicas, como cuando una proposición está incluida en otra. En estos casos, una unidad cognitiva aparece como un elemento de otra. Con ello, esta arquitectura pretende dar cuenta de ideas que resultan familiares en el campo de la psicología cognitiva bien contrastadas experimentalmente, como son la existencia y limitación de unidades cognitivas (*chunks*) en la memoria de trabajo o a corto plazo.

Dos son los aspectos a destacar respecto a la memoria declarativa en ACT*. El primero es que su granularidad es de una finura intermedia, ni tan pequeña como el término en las arquitecturas clásicas, ni tan grande como los *frames* o *schemas*. Este aspecto es relevante porque los *frames*, etc., fueron introducidos bajo el supuesto de que la unidad de organización de memoria debe ser relativamente grande con el fin de dar cuenta del carácter organizado del pensamiento humano. En ACT* esto se resuelve, como veremos, mediante la activación y la difusión de activación entre los nodos de la red, que actúa, por tanto, como mecanismo de focalización; el segundo, de la mano del anterior, es la propia estructura de la memoria, en forma de red semántica, donde cada elemento tiene una fuerza asociada, y con un mecanismo de difusión de activación.

3. *Activación*: los procesos de activación definen la memoria de trabajo, lo que implica que en ACT* esta memoria se puede considerar como un subconjunto activado de la memoria declarativa. Las producciones solamente pueden ser satisfechas por conocimiento que se en-

cuentra activado. Un aspecto a destacar es que la activación en ACT* es una propiedad de los nodos que varía de forma continua, en lugar de a intervalos discretos. Existen buenas razones de índole neurofisiológica en favor de la variación continua del nivel de activación, puesto que la activación de las neuronas (o de grupos de neuronas) parece ser continua también. El nivel de activación controla la probabilidad en la que puede ser satisfecho y su probabilidad de satisfacción exitosa. Por ejemplo, si dos estructuras pueden ser satisfechas por el mismo patrón, ACT* preferirá la más activa, cosa que es importante en casos tales como la resolución de significados ambiguos.

4. La *fuerza* en la memoria declarativa: cada nodo de la memoria declarativa cuenta con una fuerza asociada, que depende básicamente de la frecuencia de uso de la unidad cognitiva correspondiente. Esta noción de fuerza de los nodos permite asimismo definir la fuerza relativa de una asociación entre nodos. Esta propiedad es muy importante en la difusión de activación, porque tiende a transmitirse más activación sobre los caminos o vínculos con más fuerza. Además, debe tenerse en cuenta que la fuerza de un nodo determina cuánta activación puede emitir.

5. *Difusión* de activación: Una propiedad básica del concepto de difusión de activación recogido en la ecuación diferencial que propone Anderson, es que la difusión es continua, tanto en el tiempo como en la cantidad de activación. Este aspecto forma parte de la naturaleza de ACT*, como se ha dicho. De acuerdo con la ecuación, el cambio momentáneo en la activación es función del *input* hacia un nodo y de la probabilidad de decrecimiento (*decay*) espontáneo de ese nodo. El *input* hacia un nodo es una posible fuente de activación, junto con la suma de la activación proveniente de los nodos asociados ponderada por fuerzas relativas. ACT* propone, además, que el nivel de activación controla la probabilidad de satisfacción (*pattern matching*) que requiere la aplicación de producciones.

6. *Mantenimiento de la activación*: La activación es difundida a partir de los diferentes nodos fuente, cada uno de los cuales soporta un patrón de activación determinado sobre la red de la memoria declarativa. El patrón completo de activación es la suma de los patrones soportados por los nodos fuente individuales. Cuando un nodo cesa de ser una fuente, su patrón de activación decae rápidamente. Un aspecto interesante es que la activación depende también del modo como ha sido creado un nodo fuente. Si un nodo ha sido creado por la percepción de objetos del entorno y se desactiva, se puede recrear un nuevo nodo para reemplazarlo si el objeto se encuentra todavía en el foco de la percepción. En cambio, si el nodo fuente ha sido causado por una computación interna (es decir, por la acción de una producción), su activación comienza a decaer tan pronto como deja de ser usado. Este sistema de mantenimiento de la activación hace que el modelo de memoria de trabajo de

ACT* sea psicológicamente plausible en un sentido en que no lo son las arquitecturas clásicas o las que tienen estructura de listas. En ACT*, la capacidad de la memoria de trabajo se incrementa con la familiaridad en el dominio, puesto que los conceptos más familiares pueden difundir más activación. Además, ayudan a explicar la mejora de la capacidad de actuación con la práctica, ya que ésta incrementa la capacidad de la memoria de trabajo[31].

7. *Decrecimiento de la activación*: los supuestos anteriores acerca de la difusión y el mantenimiento de la activación implican asimismo el mecanismo del decrecimiento y la desaparición de la activación en la red. Cuando un nodo fuente se apaga, su patrón de activación decrece rápidamente. Este decrecimiento viene implicado por la misma ecuación diferencial de la difusión de activación.

8. *Estructura de las producciones*: Todas las producciones consisten en pares condición-acción. La condición especifica una conjunción de aspectos que tienen que ser verdaderos en la memoria declarativa. La acción especifica un conjunto de estructuras temporales para ser añadidas a la memoria.

9. *Almacenamiento de estructuras temporales*: acabamos de mencionar que las estructuras temporales pueden entrar en la memoria de trabajo mediante una de estas dos fuentes: (a) el proceso de codificación puede colocar descripciones del entorno en la memoria de trabajo; (b) la acción de las producciones puede crear estructuras para grabar los resultados de las computaciones internas. Debe observarse, además, que en ACT* el proceso de almacenamiento no opera sobre vínculos asociativos aislados, sino sobre unidades cognitivas completas (filas, imágenes y unidades proposicionales). Como ya se ha dicho, estas unidades no pueden contener más de cinco elementos; por consiguiente, existen límites acerca de cuánto puede ser codificado en un solo acto de almacenamiento.

10. *Fuerza de las producciones*: en ACT* la memoria de producciones está implementada en forma de una estructura de red similar a una red semántica, con difusión de activación, por lo que cada producción tiene una fuerza asociada. Esta fuerza se incrementa en una unidad con cada aplicación exitosa de la producción, es decir, con la práctica. Y las condiciones de las producciones con más fuerza reciben satisfacción más rápidamente.

31. En las arquitecturas clásicas se ha intentado solucionar el problema de la memoria de trabajo con técnicas de *chunking* (incluir unos items en otros de forma organizada), pero dice Anderson que estas técnicas no ayudan demasiado. La razón es que muchas tareas requieren tener disponible simultáneamente una gran cantidad de información que no puede ser razonablemente integrada como parte de un mismo *chunk*. Por ejemplo, en el *parsing* de una sentencia se necesita simultáneamente mantener información acerca de una nueva fila de símbolos, acerca del estado de cada nivel de *parsing*, acerca de la semántica de las palabras, acerca del conocimiento y las intenciones comunicativas del hablante, acerca de las referencias introducidas previamente en el curso de la conversación, etc.

11. *Selección de producciones mediante satisfacción de patrones*: La satisfacción de patrones es el mecanismo que decide qué producciones se van a aplicar. Un supuesto fundamental de todas las arquitecturas basadas en sistemas de producción es que la satisfacción de patrones subyace a todo tipo de cognición, dado su carácter de estar conducida por datos. Cuando la condición de una producción consigue una satisfacción adecuada para un conjunto de estructuras declarativas, la producción es seleccionada para su aplicación. El satisfactor de patrones es representado como una red de flujo de datos de contrastación de patrones. La probabilidad según la que estas contrastaciones son ejecutadas es una función del nivel de activación del nodo patrón que ejecuta las contrastaciones. El nivel de activación de ese nodo es una función positiva de la fuerza del nodo, del nivel de activación de las estructuras a satisfacer, y del grado de satisfacción de esas estructuras. También es una función negativa del nivel de activación de los patrones que compiten en la satisfacción de los mismos datos. Debido a que la ejecución de las contrastaciones en un nodo consume un tiempo considerable, la evidencia de una satisfacción en un nodo debe irse construyendo o decreciendo gradualmente con el tiempo. El nivel de activación del nodo, que refleja la confianza real del sistema en que el nodo debe ser satisfecho, determina la velocidad de la satisfacción de patrones. El nivel de activación de los nodos terminales de la base viene determinado por la activación de las estructuras de datos a las que están vinculados. El nivel de activación de los nodos más altos está determinado por su grado de satisfacción, por el de sus subpatrones y superpatrones y por el de los patrones alternativos. Existe un conjunto de influencias inhibitorias y excitatorias entre los nodos patrón parecido al propuesto para los sistemas conexionistas.

Efectos psicológicamente interesantes de este mecanismo son que el proceso de satisfacción de patrones permite a las producciones aplicarse aun cuando sus condiciones no hayan sido completamente satisfechas. Si la condición de una producción constituye la mejor interpretación para los datos disponibles y se encuentran disponibles suficientes partes del patrón para elevarlo por encima del umbral, la producción se aplica. Esta satisfacción parcial da cuenta de muchos de los errores en la ejecución de habilidades: por ejemplo, el uso de una palabra cercana, aunque incorrecta, en una conversación, es atribuible a una especificación incompleta de la palabra deseada en la memoria de trabajo, que es la que mejor satisface el patrón de una palabra equivocada. Por otra parte, el hecho de que la satisfacción de patrones sea tan costosa en todas las implementaciones conocidas de sistemas de producción, constituye una evidencia sugestiva en el sentido de que debiera ser el más importante cuello de botella temporal en la cognición humana, como sucede en ACT*.

12. *Procesamiento dirigido a metas*: las producciones pueden especificar una meta en su condición. Si la meta especificada satisface la

meta real, las producciones correspondientes adquieren una precedencia especial sobre las producciones que no lo hacen. Esta característica constituye una importante ventaja de la arquitectura ACT* sobre las arquitecturas clásicas, puesto que en ACT* las producciones pueden crear y transformar una estructura jerárquica de objetivos o metas que refleja el plan de acción del momento. ACT* focaliza la atención sobre una meta convirtiéndola en una fuerte fuente de activación. También puede mover el foco de atención dentro de la estructura de metas, de modo que cuando se consigue una meta se desactiva y se cambia la atención a la meta siguiente. Esta estructura de metas resuelve el problema del foco de atención de modo mucho más efectivo que los nodos de control de las arquitecturas clásicas. Su utilización da cuenta de la generalidad de la estructura jerárquica de la conducta humana, y el foco de atención que produce explica la fuerte serialidad en el flujo de la cognición humana. Una consecuencia psicológica empíricamente relevante de este esquema es el carácter serial en muchos aspectos del procesamiento cognitivo, por razón de que solamente se puede atender una meta en cada momento [32]. Tal meta se convierte en una poderosa fuente de activación, de modo que las fuentes de satisfacción de patrones son orientadas hacia la satisfacción de estructuras que contienen dicha meta.

13. *Compilación de producciones*: La compilación de conocimiento es el medio por el que entran las nuevas producciones en el sistema. En ACT* todo el conocimiento llega inicialmente en forma declarativa y tiene que ser interpretado por procedimientos generales. No obstante, con la práctica, la procedimentalización va reemplazando gradualmente la aplicación interpretativa con producciones que ejecutan la acción directamente. Este proceso de procedimentalización se complementa con un proceso de composición que puede combinar una secuencia de producciones en una sola producción. La procedimentalización y la composición, que conjuntamente equivalen a la compilación de conocimiento, crean producciones específicas de tareas mediante la práctica.

14. *Ajuste de producciones*: las producciones, una vez creadas, pueden ser ajustadas para un dominio determinado. Las producciones acumulan fuerza en relación directa a sus aplicaciones exitosas. Existen diversos procesos de aprendizaje que producen el mejoramiento gradual y continuado: los procesos de generalización y discriminación buscan aspectos del problema que son predictivos del éxito de un método particular; la composición colapsa secuencias de producciones repetidas frecuentemente en producciones únicas. Los mecanismos de aprendizaje procedimental son graduales e inductivos, y contrastan netamente con el aprendizaje directo y repentino característico del dominio declarativo.

32. Este carácter serial no impide que otros muchos procesos se computen en paralelo, la propia selección de producciones en ACT* lo hace así.

Anderson piensa que este contraste es la mejor evidencia en favor de la distinción declarativo-procedimental, que resulta básica en la arquitectura ACT* [33].

3. *La arquitectura CI (contrucción-integración)*

La arquitectura que acabamos de ver tiene un carácter general o unitario, pues a excepción de la recepción de estímulos, pretende abarcar un conjunto muy amplio y variado de actividades cognitivas superiores que van desde la comprensión del lenguaje hasta la solución de problemas de diverso tipo y la planificación. Ahora bien, de entre estas actividades, la comprensión del lenguaje ha presentado tradicionalmente problemas de modelización y tratamiento computacional. La razón es que, como ya hemos dicho, los procesos de comprensión tienen lugar en la escala de milisegundos, cosa que los había convertido en intratables mediante los esquemas tradicionales de solución de problemas en arquitecturas convencionales. Ello puede ser debido, como apunta Kintsch (1992), a que la comprensión es un proceso *sui generis* que debe ser situado en algún lugar dentro del continuo entre la percepción y la solución de problemas. Es importante insistir en este carácter continuo de muchos procesos cognitivos porque, como sugiere Kintsch, incluso podría resultar útil considerar parte de lo que hasta el momento se ha visto como solución de problemas, desde la perspectiva de la comprensión.

El modelo CI es una arquitectura para la comprensión que pretende dar cuenta de la variedad de fenómenos al que usualmente se denomina «comprensión» en el lenguaje ordinario. Según se infiere de lo que acabamos de mencionar, el éxito de esta empresa supondría demostrar que resulta útil separar los procesos de comprensión de la percepción, por una parte, de la solución de problemas, por otra. Sin embargo, tampoco se pretende afirmar que todos los casos de solución de problemas pueden ser vistos desde la perspectiva de la comprensión. CI tiene un espacio más reducido que ACT* o SOAR en el sentido de que incluso si su enfoque de la comprensión resultase tener éxito , seguiría existiendo la necesidad de una teoría propiamente dedicada a la solución de problemas, complementando, aunque no suplantando, a la teoría de la comprensión [34].

La teoría que se presenta aquí es la última versión, refinada, de la familia de modelos de comprensión de textos iniciada en Kintsch y van Dijk (1978), y van Dijk y Kintsch (1983). El mecanismo para la activa-

33. La distinción declarativo-procedimental fue objeto de una fuerte polémica hace unos pocos años, aunque actualmente no parece despertar tanto interés.

34. Puede verse, no obstante, Rosemary J. Stevenson (1992) para un enfoque integrado de la cognición favorable a la idea de que una arquitectura como CI resulta muy plausible como base para los procesos de pensamiento en la medida en que éstos descansan, en gran parte, en procesos de comprensión de lenguaje.

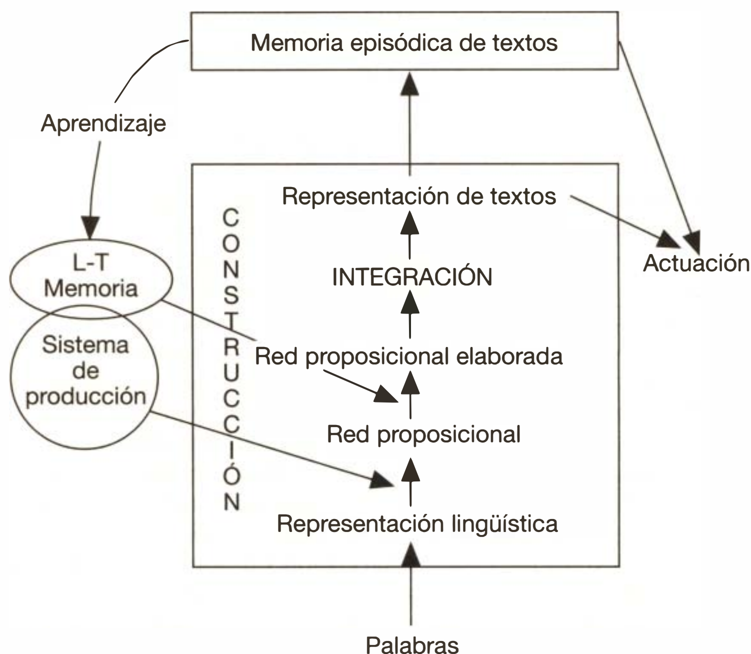


Figura 3 (tomada de Kintsch, 1992).

ción de conocimiento e inferencia fue añadido a este modelo básico de procesamiento más tarde por Kintsch (1988). El nombre «*Construction-Integration*» refleja el supuesto central acerca de la naturaleza de la interacción entre el lector y el texto en la comprensión. Desde el punto de vista arquitectónico CI es, como en cierto modo sucede con ACT*, una teoría híbrida: su componente de construcción está basado en reglas y es simbólico, pero el proceso de integración utiliza un tratamiento conexionista.

Hemos dicho que CI asume que la comprensión es un proceso interactivo, en el sentido de que siempre conlleva la interacción de un *input* externo, usualmente un texto, con los objetivos y el conocimiento del sujeto que realiza el proceso de comprensión. La figura 3 muestra un esquema de ese proceso interactivo.

Pero CI trata sólo algunos de los aspectos de esta interacción [35]. Acepta los elementos de un texto como dados y se centra en el problema

35. Por ejemplo, no trata los aspectos perceptuales implicados en la lectura. En su lugar, suele comenzar con una representación previamente codificada del texto. Kintsch piensa que quizá dentro de algún tiempo pueda utilizarse alguno de los sistemas de *parsing* actualmente en desarrollo como cabecera del sistema CI.

de cómo se combinan para formar núcleos mayores. En concreto, trata los procesos implicados en cómo activan el conocimiento las proposiciones del texto para obtener una representación integrada de texto y conocimiento. Kintsch considera, al contrario de lo que ha venido siendo habitual, que las oraciones aisladas presentan problemas de análisis más difíciles de resolver que las situadas dentro de un discurso o un contexto más general, y así es como viene caracterizada la comprensión en CI. Por consiguiente, las representaciones generadas por sus procesos son estructuras holistas. Estos procesos responden al siguiente esquema: se construye una red asociativa de elementos (conceptos, proposiciones...) para representar el significado de un discurso o situación como un todo, pero al mismo tiempo los elementos derivan parte de su significado a partir de la estructura en la que están insertos. El funcionamiento de la red está basado en principios de satisfacción de restricciones, es decir, los elementos se afectan entre sí positiva o negativamente hasta que la red alcanza un estado de equilibrio que consiste en la dominancia del núcleo de la red interrelacionado positivamente después de rechazar la parte marginal.

Es importante resaltar el hecho de que la activación de conocimiento en CI es local, asociativa, no se guía por el control de estructuras organizadas como los *frames* o *schemas*, y por tanto resulta primitiva e imprecisa. Ello se debe a que se activa tanto conocimiento relevante como irrelevante. Posteriormente, el proceso de integración de tipo conexionista es capaz de proporcionar una representación coherente del texto rechazando el conocimiento contextualmente irrelevante y el material contradictorio. El resultado de este proceso es una representación de texto que puede ser base para la acción. La representación se almacena en la memoria episódica de texto (responsable de tareas como el recuerdo y la sumarización), y finalmente puede modificar la memoria a largo plazo, con lo que se puede afirmar que CI es una arquitecura para el aprendizaje en sentido estricto.

3.1. El proceso de construcción

Es bien conocido que la estructura de un texto viene determinada por lo que está escrito y por lo que hace el lector. Lo que aporta CI es un mecanismo computacional preciso que permite describir la naturaleza y

36. En este caso, por ejemplo, los vínculos causales han sido dotados con más fuerza que los de repetición de argumento (Mannes & Kintsch, 1991). Las hipótesis mutamente excluyentes son conectadas con vínculos inhibitorios, y en los casos de asimetría, como cuando un procedimiento contiene una precondición para otro, también se establecen vínculos inhibitorios. Piénsese, por ejemplo, en una de las aplicaciones de CI como el proceso de comprensión del sistema UNIX, donde el procedimiento *edit file* inhibe al procedimiento *delete file*, ya que este último podría eliminar una precondición para el primero, aunque no al revés.

los resultados de esta interacción. Los aspectos a destacar en el proceso de construcción de CI son los siguientes. El primero, que un *parser* para CI no necesita ser perfecto. Si construye formas menos apropiadas en adición a las apropiadas, el contexto funciona de forma suficientemente restrictiva como para eliminar las equivocadas. Es decir, el mismo mecanismo conexionista que permite seleccionar el conocimiento apropiado a partir de todo el conocimiento activado puede ser usado para simplificar el *parsing*.

Otra de las propiedades más interesantes de CI es su mecanismo de *focalización*. En el proceso de comprensión, además de formar conceptos y proposiciones, es preciso determinar sus relaciones. Pero este es un proceso difícil de diseñar, puesto que si el texto es de un tamaño razonable, no resulta posible (ni tampoco resultaría plausible psicológicamente) mantener en la memoria de trabajo todas las proposiciones del texto. Por otra parte, los textos son leídos oración por oración, de modo que sólo pueden procesarse en la memoria de trabajo unidades elementales una por una. En esta situación, ¿cómo resulta posible construir una representación coherente del texto global? CI asume que cuando el lector pasa de una unidad a otra no se desactiva todo el material de la memoria de trabajo, sino que queda retenido un número reducido de proposiciones (entre una y cuatro, con un promedio de dos para el lector normal). Estas proposiciones son mantenidas en el foco de atención y reprocesadas junto con el material perteneciente al siguiente ciclo de procesamiento. Si todo va bien, esas proposiciones retenidas forman un vínculo entre la parte del texto procesada previamente y el nuevo *input*, con la particularidad de que en CI las proposiciones más altamente activadas cuentan con mayor probabilidad de servir de puente. Esto hace que, en la mayoría de los casos, las proposiciones de la memoria a largo plazo más altamente asociadas a las proposiciones del texto resultarán activadas e incorporadas a la representación del texto en construcción. Por otra parte, los objetivos del lector desempeñan un papel relevante en la activación de conocimiento, incluido el procedimental, relacionado con esas metas.

Con los ingredientes anteriores se genera una representación de texto que corresponde a la red proposicional elaborada de la figura. Esta representación contiene: *a)* un conjunto de proposiciones construidas a partir del texto; *b)* unas pocas proposiciones asociadas a cada proposición del texto recuperadas de la memoria a largo plazo en función de la fuerza de la asociación; *c)* objetivos específicos asociando conocimiento declarativo y procedimental. De todos modos, contiene muchos *items* irrelevantes, procedimientos que no se necesitan para la tarea a realizar, e incluso *items* contradictorios (por ejemplo, hipótesis alternativas). Para resolver esta situación se necesita el proceso de integración.

3.2. El proceso de integración

Este proceso viene determinado por la estructura de la representación elaborada. Como ya se ha anticipado, esta representación forma una red ricamente interrelacionada de proposiciones. Las relaciones pueden ser de diversos tipos, por ejemplo, compartir el mismo referente (repetición de argumento) o relaciones semánticas más específicas como la causalidad, y varían en fuerza dependiendo de su tipo. Además, sus vínculos no siempre son positivos ni simétricos. La matriz que especifica todos los vínculos en la representación elaborada de texto se denomina matriz de coherencia. Por lo ya dicho, el mecanismo de integración es fácil de comprender. Inicialmente todas las proposiciones que provienen del texto son dotadas con una fuerza de 1, y las basadas en el conocimiento con una fuerza de 0. El procedimiento de difusión de activación entonces se encarga de que la red alcance un estado de equilibrio. Si todo sucede con normalidad, los grupos de proposiciones más fuertemente interconectados acaparan la mayor parte de la activación de la red, desactivando las partes aisladas y los nodos con vínculos negativos[37].

Este proceso de integración tiene lugar en un único ciclo *input*, y su resultado representa el contenido de la memoria de trabajo al final del ciclo. Cuando el proceso se mueve al ciclo siguiente, suceden dos cosas. En primer lugar, la estructura global generada es eliminada de la memoria de trabajo y almacenada en la memoria secundaria como una representación episódica de texto (ver figura). En segundo lugar, las proposiciones más altamente activadas permanecen en la memoria de trabajo durante el procesamiento de la siguiente secuencia. Es decir, al final de cada ciclo, el foco de atención se mueve hacia un nuevo *input*, a excepción de las proposiciones más relevantes del ciclo anterior, que continúan activadas[38]. La memoria episódica de texto consiste precisamente en los resultados acumulados en cada ciclo de procesamiento.

Anteriormente hemos hecho alusión al aprendizaje en CI. Este se define en términos de los efectos que produce la memoria episódica de textos en la memoria a largo plazo. No obstante, Kintsch piensa que está por averiguar cómo se conceptualiza este proceso, aunque ofrece dos posibilidades. Una es que la memoria episódica simplemente se combina

37. Desde el punto de vista matemático, esta operación consiste en multiplicar el vector inicial de activación por la matriz de coherencia, y a repetir esta operación hasta que el patrón de activación se estabiliza.

38. Esta propiedad focalizadora de la arquitectura CI está muy apoyada desde la propia psicolingüística, como un requisito imprescindible en el proceso de comprensión del discurso para resolver, por ejemplo, referencias anafóricas. Puede verse al respecto Grosz (1981) y Grosz & Sidner (1990). Y para una propuesta de incoporar un mecanismo de focalización a una arquitectura de procesamiento para el razonamiento reflejo como la de L. Shastri (1993) con objeto de capacitarlo para la resolución de referencias anafóricas, puede verse Ezquerro & Iza (1993).

con la red de la memoria a largo plazo, y la otra que los vínculos de la red de la memoria a largo plazo se actualizan con la información que proviene de la memoria episódica.

CI ha sido aplicada a una variedad de tareas cognitivas como la identificación de palabras en un contexto y el reconocimiento de sentencias, tratando de comparar sus resultados con los de los experimentos conocidos. Una de las pruebas más interesantes a que se puede someter a una teoría de la arquitectura es ver cómo funciona con tareas para las que no ha sido expresamente diseñada. Esto es lo que hicieron Kintsch (1991) y Kintsch & Welsch (1991). Después de tomar resultados experimentales clásicos sobre memoria, los simularon en CI con el fin de explorar si la teoría general era lo suficientemente rica como para permitir la derivación de modelos adecuados para esos fenómenos sin tener que acudir a supuestos *ad hoc*. Los recursos de CI demostraron ser suficientes para dar cuenta del papel de las señales sintácticas en un texto indicando la relevancia del discurso, un análisis del papel desempeñado por las relaciones causales en la comprensión de historias y algunos resultados de *priming* en contextos de discurso. CI ha sido aplicada a la comprensión de problemas aritméticos (Kintsch & Greeno, 1985), y a la simulación de los procesos de comprensión de las instrucciones para la ejecución de tareas rutinarias de computación (Mannes & Kintsch, 1991).

VI. EL PROBLEMA DE LA EVALUACIÓN DE LAS ARQUITECTURAS

Evaluar empíricamente una arquitectura no es tarea fácil, y sin embargo es algo que, inevitablemente, hay que tratar de hacer. La razón obvia es que, como hemos dicho, una teoría de la arquitectura es una teoría acerca de los mecanismos que soportan la actividad inteligente. Como tal establece las restricciones fundamentales de la cognición, y por tanto, es la principal portadora de los compromisos empíricos de una teoría cognitiva. Antes hemos expuesto algunos recelos filosóficos ante la posibilidad de las arquitecturas, y también ha sido aludida la tesis de la indeterminación de la traducción de Quine, en el sentido de que parece imponer restricciones adicionales a la posibilidad de corroboración empírica de las teorías psicológicas. Pero debe tenerse en cuenta que esta tesis de Quine afecta al problema de la determinación de los contenidos mentales, no a la arquitectura en sentido estricto. Es conveniente tener esto en cuenta, porque muchos argumentos filosóficos en contra de las posibilidades de la inteligencia artificial para modelar la mente descansan en el confuso supuesto de que una teoría cognitiva ideal debería modelar todos los aspectos internos responsables de la conducta, tanto los debidos al conocimiento como los arquitectónicos. A estas alturas debiera ser obvio

que una misma conducta abierta puede ser resultado de una inmensa variedad de procesos cognitivos internos. En la medida en que no existen dos organismos cognitivos idénticos en lo que respecta a su base de conocimiento, ya que todos han tenido historias de aprendizaje diferentes, es inútil pretender hallar un modelo de simulación que sea fuertemente equivalente a cualquier conducta. Pero esta limitación pertenece a la propia naturaleza de los sistemas de procesamiento de información, y debiera ser suficiente para mostrar la inutilidad y trivialidad de los argumentos filosóficos de imposibilidad *a priori* basados en ella. Una teoría de la arquitectura debe tratar de caracterizar, como dice Pylyshyn, un «punto fijo cognitivo», común a todos los organismos de la misma especie. Si se diera el caso de que la arquitectura cambiase en formas que requieren una explicación cognitiva basada en conocimiento (reglas y representaciones), entonces la arquitectura no podría ser usada para explicar cómo los cambios en el conocimiento dan lugar a cambios de conducta. No obstante, las dificultades para distinguir netamente entre la arquitectura y el conocimiento podrían abrir de forma indirecta la puerta a este tipo de argumentos. Pero debe quedar claro que, en principio, se trata de problemas independientes.

Acabamos de decir que la contrastación de teorías psicológicas no es tarea fácil. Si nos planteamos la cuestión de si un modelo psicológico corresponde o no a un sistema procesador de información, nos toparemos con el problema de que la noción de correspondencia es poco clara, ya que, aplicada a un sistema cognitivo, puede darse a muchos niveles. De ahí que con frecuencia la contrastación en psicología se haya limitado a comprobar si el modelo realiza la misma función *input-output* que el sistema u organismo modelado, pero ya hemos visto que este tipo de contrastación deja a las teorías psicológicas casi completamente indeterminadas. Un paso más es tratar de comprobar si el modelo realiza la función del mismo modo que el sistema modelado. Sucede igualmente que la noción de «modo» o «método» tampoco está suficientemente definida, pero ofrece, no obstante, mayores pistas para avanzar. Quizá el análisis de protocolos anteriormente expuesto como introducción a los sistemas de producción podría situarse en este punto. Este método, no obstante, resulta bastante limitado al ser aplicable solamente a un número reducido de tareas, en general, casos de solución de problemas relativamente lentos y deliberados. Aun así, este análisis proporciona evidencia acerca de estados intermedios que de otra forma no estarían disponibles, y su importancia empírica puede incrementarse combinándolo con otras técnicas independientes como grabaciones en vídeo de los gestos de los sujetos, medición de los movimientos de los ojos, etc. La búsqueda de evidencia acerca de estados intermedios no sólo sirve para ofrecer pistas acerca del «cómo» se ejecuta una tarea, también resulta útil para deslindar los elementos del proceso debidos a la arquitectura y los

achacables al conocimiento. Como dice Pylyshyn (1989), cuando hay evidencia de que un estado intermedio de un proceso es transparente, o resulta accesible a otra parte del sistema, puede ser considerada asimismo como evidencia de que tal proceso no es primitivo, sino que admite descomposición ulterior.

Otra forma de hablar del «modo» a un mayor nivel de profundidad consiste en tratar de especificar en detalle la secuencia de pasos que atraviesa un sistema cuando ejecuta una función. Hacer esto supone proporcionar un algoritmo para el proceso en cuestión, y esta noción sí que está mejor definida. Antes hemos dicho que algo fundamental y básico que hace la arquitectura es determinar la clase de algoritmos que son ejecutables *directamente* en ella, en lugar de ser sólo simulables previa emulación de otra arquitectura. Si este supuesto es correcto, entonces la búsqueda de metodologías para evaluar la equivalencia fuerte debe proporcionar evidencia en favor de la arquitectura. Uno de los autores que más se ha preocupado de reflexionar acerca de las metodologías para evaluar la equivalencia fuerte es Z. Pylyshyn[39]. El criterio básico que propone para decidir cuándo estamos hablando de la arquitectura es el de *impenetrabilidad* cognitiva. Un proceso de un sistema u organismo es cognitivamente impenetrable cuando su funcionamiento no depende del conocimiento del organismo en cuestión. Si, por el contrario, la conducta del sistema es sensible a sus creencias, deseos, etc., se trata de un proceso cognitivamente penetrable. Este criterio, como se ve, tiene consecuencias empíricas, pues permite diseñar dispositivos experimentales para determinar si ciertos fenómenos psicológicos pueden ser alterados sistemáticamente cambiando el conocimiento del organismo o no.

Otro método es la búsqueda de evidencia para evaluar la equivalencia de complejidad. Este criterio proviene de consideraciones computacionales y está asociado con el uso de mediciones de tiempos de reacción y medidas de demanda de atención en tareas. Su fiabilidad experimental se ha incrementado con el tiempo, pues hasta hace pocos años no existían técnicas como la cronometría mental. La importancia de este método se basa en la idea de que la relación entre el número de pasos primitivos empleados y ciertas propiedades del *input* se puede considerar como una propiedad invariante esencial de lo que intuitivamente se piensa que son realizaciones diferentes de un mismo algoritmo. Por ejemplo, no se pueden considerar dos procesos como realizaciones de un mismo algoritmo si uno de ellos computa una función en un tiempo fijo, con independencia del tamaño del *input*, mientras que el otro incrementa exponencialmente el tiempo empleado a medida que varía alguna propiedad del *input*. Lo que importa aquí es la naturaleza de la relación entre factores como el tiempo consumido, el número de pasos y las

39. Véase Pylyshyn (1978, 1984 y 1989).

propiedades del *input*. Como hemos dicho, el método común utilizado en psicología cognitiva para evaluar la complejidad relativa es la medición de los tiempos relativos de reacción, estableciendo relaciones entre el tiempo empleado para una tarea y determinadas propiedades paramétricas de la tarea como, por ejemplo, el tamaño del input. Es razonable pensar que el conjunto de procesos que son equivalentes en cuanto a complejidad representa un refinamiento con respecto al conjunto de procesos que computan la misma función *input-output*. De cualquier modo, la equivalencia de complejidad no puede ser por sí misma suficiente para definir la equivalencia fuerte, aunque puede ser una condición necesaria.

Además de estos criterios generales, en la práctica aparecen problemas concretos. Kintsch (1992) ha señalado algunos problemas a partir de su experiencia de contrastación de CI. Uno de ellos está relacionado con el hecho de que, típicamente, los ejemplos que se usan para simular las arquitecturas suelen ser los mismos que previamente usan los experimentadores originales para introducir y discutir sus datos. Estas prácticas son cuestionables, pues descansan absolutamente en la tipicalidad de los ejemplos, y un ejemplo típico no es lo mismo que los datos promedio sobre sujetos. Sin embargo, los datos promedio tienen sus problemas también, ya que definen sujetos ideales y descuidan las diferencias individuales, en la medida en que la variabilidad es tratada como ruido y no como diferencia real. El remedio a estos problemas consiste en combinar los dos procedimientos. Es decir, simular ejemplos propotípicos y a la vez comparar los resultados de la simulación con datos promedio extraídos a partir de muchos ejemplos y con los obtenidos de la contrastación con sujetos individuales.

Otra de las dificultades que debe afrontar una arquitectura general es el hecho de que prácticamente ninguna aplicación particular de la teoría utiliza todos sus recursos. Por ejemplo, en el estudio del reconocimiento de sentencias, el proceso de integración de CI desempeña un papel central, pero no lo hacen ni la naturaleza cíclica del proceso ni tampoco la activación de conocimiento. Otro tanto sucede con el componente de la activación cuando se ha aplicado CI a la comprensión de instrucciones para ejecutar tareas rutinarias de computación. Lo peculiar de este caso es que, aunque al principio el hecho de que cada proposición del texto active algunas proposiciones relacionadas asociativamente desempeña un papel despreciable, pasa a tener un papel dominante una vez que el sistema ha ejecutado unas pocas tareas y recuerda lo que ha hecho. Casos como éste ponen de manifiesto la importancia de la búsqueda de predicciones cualitativas y no solamente cuantitativas. De cualquier modo, el problema de la contrastación fragmentaria no debería asustar, puesto que es una consecuencia natural de aplicar una arquitectura que cubre un rango muy amplio de fenómenos. Cada tarea cognitiva hace uso de

una porción limitada de las capacidades de una arquitectura de este tipo, pero parece razonable pensar que lo mismo sucede con la gente real. Lo interesante del caso es precisamente que dentro de una misma arquitectura se pueda dar cuenta de una gran variedad de habilidades cognitivas.

Las técnicas para la evaluación de las arquitecturas irán ganando fiabilidad a medida que vayamos esclareciendo y delimitando el propio papel de la arquitectura y su utilidad. Newell y col. (1989) sugieren cuatro posibles respuestas a esta cuestión: *a) Establecimiento de parámetros generales*: aunque desde esta perspectiva la arquitectura tiene efectos de largo alcance en el sentido de que interviene en todo el procesamiento cognitivo, sus efectos se pueden resumir en unos pocos parámetros generales. Entre ellos están el tiempo de una operación elemental, el tamaño de la memoria a corto plazo, la tasa de adquisición o aprendizaje en la memoria a largo plazo, el tiempo consumido en realizar un movimiento en el espacio de un problema, etc.; *b) Determinación de la conducta cognitiva simple*: ya hemos dicho anteriormente que a medida que se va reduciendo el tiempo para la ejecución de una tarea, se van reduciendo igualmente las opciones en cuanto a las posibles secuencias a seguir. Esta idea puede ser usada de dos formas. En primer lugar, para diseñar técnicas experimentales que proporcionen evidencia acerca del tiempo aproximado de ejecución de una secuencia elemental. En segundo lugar, si partimos de una arquitectura dada, y por tanto con el tiempo por secuencia elemental establecido, su comparación con la conducta real tiene consecuencias empíricas en el sentido de que los resultados de la comparación nos pueden indicar si la tarea es realizable por la arquitectura en cuestión o no, y si lo es, de qué modo. Por ejemplo, dados los parámetros de ACT*, hay que descartar que determinadas tareas sean efecto de activaciones múltiples de producciones, debido al consumo de tiempo tan elevado que necesitan las producciones para activarse en ACT* (y en cualquier arquitectura basada en producciones), por tanto, tienen que deberse a fenómenos de paralelismo y difusión de activación; *c) Conexiones ocultas*: las teorías de la arquitectura brindan, como hemos mantenido a lo largo de este escrito, una forma de unificación para la ciencia cognitiva, pues asumen que todos los humanos realizamos nuestras actividades inteligentes mediante el mismo conjunto de mecanismos, a pesar de que el conocimiento cambie de persona a persona. En este contexto, un resultado importante de la investigación en teorías de la arquitectura debe ser el sacar a la luz conexiones ocultas entre actividades cognitivas que, por otra parte, pueden parecer muy distintas sobre la base del conocimiento y la situación[40]; *d) Despejar grados de libertad*: un

40. Un ejemplo de resultados en este contexto puede ser el papel central que ha adquirido el *chunking* en muchas formas de aprendizaje diferentes.

hecho bien conocido desde los comienzos de los intentos de simular la cognición que ha causado muchos dolores de cabeza, es el de que para conseguir que una simulación funcione es necesario especificar muchos procedimientos y estructuras de datos que no tienen ninguna justificación psicológica. El hecho se agrava porque la estructura de los programas no ofrece ninguna pista acerca de qué aspectos suponen hipótesis psicológicas y cuáles no. Sin embargo, el trabajo en teorías de la arquitectura puede ayudar a remediar en parte este problema. La razón es que una propuesta de arquitectura debe ser una propuesta de un sistema operacional completo. Por consiguiente, cuando se hace una simulación dentro de una arquitectura, todos los aspectos del sistema representan hipótesis empíricas, y en consecuencia, su plausibilidad cognitiva depende asimismo de la contrastación empírica.

Como se podrá ver, ninguna de estas técnicas es, por sí sola, capaz de proporcionar criterios suficientes para la determinación de la equivalencia fuerte, y por consiguiente, tampoco desvelan del todo la arquitectura. Pero también muestran, contrariamente a lo que muchos argumentos un tanto simplistas han tendido a asumir, que la mera conducta observable no es todo lo que tenemos en ciencia cognitiva. Si fuera así, no habría manera de ir más allá de la equivalencia débil estímulo-respuesta para comparar modelos cognitivos. Y no parece ser el caso. Seguramente, buena parte de los aspectos de las arquitecturas que acabamos de ver son falsos, pero, como dice Newell, ese es simplemente el destino de las teorías erróneas que no corresponden a la realidad. En este sentido, la investigación en arquitecturas y la simulación de procesos cognitivos en ellas es lo que puede hacer de la ciencia cognitiva una disciplina empírica, más allá de la pura inteligencia artificial y de la simple recogida y elaboración estadística de datos, o correlaciones estímulo-respuesta, sin saber qué teorías apoyan o refutan, simplemente porque no se aventuran teorías.

BIBLIOGRAFÍA

Anderson, J. (1983), *The Architecture of Cognition*, Harvard University Press, Cambridge, Mass.

Anderson (1990), *The adaptive character of thought*, Erlbaum, Hillsdale, NJ.

Anderson (1991), «Is human cognition adaptive»: *Behavioral and Brain Sciences*, *14*, 471-517.

Anderson (1991a), «The place of cognitive architectures in a rational analysis», en K. van Lehn (ed.), *Architectures for Intelligence*, Erlbaum, NJ.

Charniak, E. y McDermott, D. (1985), *Introduction to Artificial Intelligence*, Addison-Wesley, Reading, Mass.

Dennett, D. (1988), «When Philosophers Encounter Artificial Intelligence»: *Daedalus, 117*, 283-296.

Dennett, D. (1991), *Consciousness Explained*, Little, Brown and Co., Boston.
Dennett, D. (1994), «The Practical Requirements for Making a Conscious Robot»: *Philosophical Transactions of the Royal Society* (en prensa).
Ezquerro, J. (199_), «Acciones, planes y tecnología, en F. Broncano» (comp.), *Nuevas meditaciones sobre la técnica*, Trotta, Madrid (en prensa).
Ezquerro, J. y Iza, M. (1993), *Reflexive Reasoning, Focus Theory and Anaphora Resolution*, Logic Seminar Report LPHS-EHU-02.3, Departamento de Lógica y Filosofía de la Ciencia, Universidad del País Vasco.
Feldman, J. A. (1985), «Connectionist models and their applications. Introduction»: *Cognitive Science*, *9*, 1-2.
Fodor, J. (1974), «Special Sciences: or the disunity of Sciences as a Working Hypotheses»: *Synthese*, *28*, 77-115.
Fodor, J. (1983), *The Modularity of Mind*, MIT Press, Cambridge, Mass.
Fodor, J. (1987), «Modules, Frames, Fridgeons, Sleeping Dogs and the Music of Spheres», en I. L. Garfield, 1987.
Fodor, J. y Pylyshyn, Z. (1988), «Connectionism And Cognitive Architecture: A Critical Analysis»: *Cognition*, *28*, 3-71.
Galton, A. (1993), «On the Notions of Specification and Implementation»: *Philosophy*, Supplement, *34*, 111-136.
Garfield, I. L. (1987), *Modularity in Knowledge Representation and Natural Language Understanding*, MIT, Bradford.
Grosz, B. J. (1981), «Focusing and description in natural languages dialogues», en A. Joshi, B. Webber y I. Sag (eds.), *Elements of discourse understanding*, Cambridge University Press, Cambridge.
Grosz, B. J. y Sidner, C. L. (1990), «Plans for Discourse», en P. R. Cohen, J. Morgan y M. E. Pollack (eds.), *Intentions in Communication*, MIT Press, Cambridge Mass.
Hayes, P. J. (1987), «What the Frame Problem Is and Isn't», en Z. Pylyshyn (ed.), 1987, 123-137.
Horgan, T. y Tienson, J. (1993), «Levels of Description in Nonclassical Cognitive Science»: *Philosophy*, Suplement 34, 159-188.
Kintsch y van Dijk (1978), «Towards a Model of Text Comprehension and Production»: *Psychological Review*, *85*, 363-394.
Kintsch (1988), «The use of knowledge in discourse processing: a construction-integration model»: *Psychological Review*, *95*, 163-182.
Kintsch (1991), «How readers construct situation models for stories. The role of syntactic cues and causal inferences»: en A. F. Healy, S. M. Kosslyn y R. M. Shiffrin (eds.), *From learning processes to cognitive processes: Essays in honour of William K. Estes* (vol. 2). Erlbaum, Hillsdale, NJ.
Kintsch, W. y Greeno, J. G. (1985), «Understanding and solving arithmetic problems»: *Psychological Review*, *92*, 109-129.
Kintsch y Welsch (1991), «The construction-integration model: A framework for studying memory for text», en W. E. Hockley y S. Lewandowsky (eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock*, 367-385, Erlbaum, Hillsdale, NJ.
Kintsch, W. (1992), «A cognitive architecture for comprehension», en H. L. Pick Jr., P. van der Broek and D. C. Knill (eds.), *Cognition, Conceptual and Methodological Issues,* American Psycological Association, Washington.

Laird, J. E., Newell, A. y Rosembloom, P. S. (1987), «SOAR: An Architecture for General Intelligence»: *Artificial Intelligence*, *33*(1), 1-64.
Lynn Foster, C. (1990), *Algorithms, Abstraction and Implementation. A Massively Multilevel Theory of Strong Equivalence of Complex Systems*, Ph. D. Thesis, University of Edinburgh.
Mannes y Kintsch (1991), «Planning routine computing tasks: Understanding what to do»: *Cognitive Science*, *15*, 305-342.
Marr, D. (1982), *Vision*, Freeman, San Francisco.
Mcclamrock, R. (1991), «Marr's Three Levels: A Re-Evaluation»: *Minds And Machines*, *1*, 185-196.
McClelland, J. y Rumelhart, D. (1985), «Distributed memory and the Representation of General and Specific Information»: *Journal of Experimental Psychology*, *144*, 159-188.
Martí-Oliet, N. y Meseguer, J. (1993), *Action Change in Rewriting Logic*, Tecnical Report, Computing Science Laboratory, SRI International, Menlo Park, Cal.
Miller, G. A., Galanter, E. y Pribram, K. H. (1960), *Plans and the Structure of Behaviour*, Holt, New York.
Newell, A., Shaw, J. C. y Simon, H. (1958), «Elements of a Theory of Human Problem-Solving»: *Psychological Review*, *65*, 151-166
Newell, A., Shaw, J. C. y Simon, H. (1963), «Chess Playing Programs and the Problem of Complexity», en E. A. Feigenbaum y J. Feldman (eds.), *Computers and Thought*, McGraw Hill, New York.
Newell, A., Rosenbloom, P. S. y Laird, J. E. (1989), «Symbolic Architectures for Cognition», en Posner (ed.), 1989, 93-131.
Newell, A. (1990), *Unified Theories of Cognition, Harvard University Press,* Cambridge, Mass.
Oaksford, M.R., Chater, N. J. y Stenning, K. (1990), «Connectionism, Classical Cognitive Science and Experimental Psychology»: *AI and Society*, *4*, 73-90
Putnam, H. (1988), *Representations and Reality*, MIT Press, Cambridge, Mass.
Pylyshyn, Z. (1978), «Computational Models and Empirical Constraints»: *Behavioral and Brain Sciences*, *1*, 93-99.
Pylyshyn, Z. (1984), *Computation and Cognition*, MIT Press, Cambridge, Mass.
Pylyshyn, Z. (ed.) (1987), *The Robot's Dilemma: The Frame Problem in AI*, Ablex Publ, N. Jersey.
Pylyshyn, Z. (1989), «Computing in Cognitive Science», en M. I. Posner, *Foundations of Cognitive Science*, MIT Press, Cambridge, Mass., 51-91.
Rumelhart, D. (1989), «The Architecture of Mind: A Connectionist Approach», en M. I. Posner, *Foundations of Cognitive Science*, MIT Press, Cambridge, Mass., 133-59.
Schneider, W. y Shiffrin, R. M. (1977), «Controlled and automatic human information processing: I. Detection, search and attention»: *Psychological Review, 84*, 1-66
Searle, J. (1980), «Minds, Brains and Programs»: *The Behavioral and Brain Sciences*, *3*, 417-424.
Sharples, M. et al. (1989), *Computers and Thought. A practical Introduction to Artificial Intelligence*, MIT, Bradford, Mass.
Shastri, L. (1990), «Connectionism and the computational effectiveness of reasoning»: *Theoretical Linguistics*, *16*(1), 65-87.

Shastri, L. y Ajjanagadde, V. (1993), «From simple associations to systematic reasoning: A connectionist representation of rules, variables, and dynamic bindings using temporal synchrony»: *Behavioral and Brain Sciences*, *16*(4).

Shiffrin, R. M. y Schneider, W. (1977), «Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory»: *Psychological Review*, *84*, 127-190.

Shoham,Y. (1987), «What is the Frame Problem?», en Georgeff y Lansky (eds.), *Reasoning about actions and plans*, Morgan Kaufman Publ., Los Altos, California.

Stevenson, R. J. (1992), *Language, Thought and Representation*, John Wiley & Sons, Chichester.

Swartout, W. y Balzer, R. (1983), «On The Inevitable Interwining Of Specification and Implementation»: *Communications Of The Acm*, *25*(7), 438-440.

Van Dijk y Kintsch (1983), *Strategies of discourse comprehension*, Academic Press, San Diego, CA.

Winograd, T. y Flores, F. (1986), *Understanding Computers and Cognition*, Ablex Publ. Co., New York.

EL CONEXIONISMO Y SU IMPACTO EN LA FILOSOFÍA DE LA MENTE

Josep E. Corbí y Josep L. Prades

I. INTRODUCCIÓN

En sus inicios [1], a principios de los cuarenta, la investigación en Inteligencia Artificial responde a dos modelos fundamentales, a saber: el simbólico y el neural o conexionista. Por un lado, los modelos simbólicos intentan simular las capacidades cognitivas de los seres humanos mediante el procesamiento de fórmulas sintácticas, pues se entiende que tales capacidades descansan necesariamente en un sistema de representación cuyos elementos básicos reflejan la estructura sintáctica del lenguaje. Por otro lado, los modelos neurales o conexionistas encuentran en la estructura sináptica del cerebro la clave para simular la inteligencia humana y, por ello, tratan de imitar la actividad inteligente a partir de la elaboración de redes de conexiones entre unidades muy simples. La aparición en 1969 del libro de Marvin Minsky y Seymour Papert, *Perceptrons*, alteró este equilibrio al cuestionar la viabilidad de la estrategia conexionista.

Minsky y Papert argumentaron que los modelos neurales difícilmente podrían llegar a simular algún día habilidades cognitivas mínimamente complejas. Esta conclusión provocó al abandono de la rama conexionista de la Inteligencia Artificial, con lo que, desde entonces, todos los esfuerzos se centraron en los modelos simbólicos. Esta opción venía, además, avalada por datos procedentes de otras disciplinas como la lingüística, donde Noam Chomsky [2] reivindicaba la existencia de una sin-

1. Entre los trabajos más significativos de este periodo inicial se cuentan McCulloch y Pitts (1943), Hebb (1949), Neuman (1956), Rosenblatt (1959, 1962) y Selfridge (1959).
2. Cf. Chomsky (1957, 1959, 1968)

taxis universal e innata para dar cuenta de la sorprendente productividad de nuestra actuación lingüística. El impacto filosófico de esta maniobra excluyente en favor del paradigma simbólico fue más que notable, y dio lugar a lo que a lo largo de este capítulo denominaremos la *Imagen Sintáctica de la Mente*[3].

Con todo, el acuerdo en torno a la exclusividad de los modelos simbólicos se quebró a principios de los años 80; pues, por una parte, los modelos simbólicos habían tropezado con dificultades importantes que empezaban a ahogar el entusiasmo inicial[4] y, por otra, surgían nuevos y atractivos modelos conexionistas que cosechaban éxitos donde los modelos simbólicos parecían fracasar[5]. Ante el reverdecer de los modelos neurales, el paradigma simbólico ya no puede dar por sentado su dominio y la Imagen Sintáctica de la Mente inicia su defensa frente al acoso de la concepción de la mente asociada a la estrategia conexionista. Así pues, parece que la mejor manera de entender el impacto del conexionismo en la filosofía de la mente consiste en ver en qué medida la Imagen Sintáctica se ve afectada por el surgimiento de los modelos neurales.

En la sección II, examinaremos el contexto en el que el desarrollo de la Inteligencia Artificial parece resultar relevante para nuestra concepción de lo mental. En este sentido, veremos que, hasta mediados de los ochenta, se reivindica la Imagen Sintáctica de la Mente como el único modo de hacer inteligible la relación entre los contenidos mentales de un organismo y las propiedades físicas de su cuerpo. Dedicaremos, por tanto, las secciones III y IV a exponer los elementos centrales de la Imagen Sintáctica de la Mente, así como a indicar la naturaleza de los argumentos que la apoyan y las dificultades con las que tropieza. En las secciones V y VI se explican brevemente las peculiaridades de los modelos conexionistas y se exploran sus consecuencias para los distintos rasgos de la Imagen Sintáctica, con lo que acabará perfilándose una nueva imagen de lo mental, a saber: la Imagen Conexionista. Esta nueva Imagen nos llevará a revisar, en la sección VII, el concepto clásico de teoría cognitiva, así como a iluminar el debate en torno al holismo de lo mental. En la sección VIII se concluye recapitulando los principales puntos y formulando algunas reservas.

3. Jerry Fodor y Zenon Pylyshyn son los representantes más significativos de la lo que aquí denominaremos «Imagen Sintáctica de la Mente»: cf. Fodor (1975, 1983, 1987, 1990) y Pylyshyn (1984).

4. Cf., en este sentido, Dreyfus (1979), Feldmand and Ballard (1982), Searle (1980, 1984, 1990), Dennett (1983, 1984), Churchland y Churchland (1990).

5. Hinton y Anderson (1981) y Rumelhart y McLelland (1986) son dos textos centrales en este punto.

II. FISICALISMO, CONTENIDO MENTAL E INTELIGENCIA ARTIFICIAL

Los programas de investigación en Inteligencia Artificial están, en principio, destinados a construir máquinas que simulen las capacidades cognitivas de los seres humanos. Se entiende, con todo, que el desarrollo de tales programas no sólo incrementará nuestra habilidad tecnológica, sino que permitirá aumentar nuestra comprensión de la estructura de la inteligencia, a la vez que mostrar cómo un conjunto debidamente organizado de materiales puede llegar a pensar. La mejora en el conocimiento de nuestras estructuras mentales favorecería el desarrollo de la psicología, mientras que el análisis de cómo ciertas estructuras físicas pueden tener propiedades mentales nos haría avanzar en la resolución del problema ontológico de la relación entre la mente y el cuerpo. Entendemos, precisamente, que la Inteligencia Artificial alcanzará su mayor relevancia en el ámbito de la filosofía de la mente en la medida en que contribuya a la elucidación del mencionado problema ontológico. De tal esfuerzo se derivan otras consecuencias acerca del contenido mental, de la articulación de las teorías científicas, etc., que se mencionarán en la penúltima sección de este capítulo.

El punto de partida de la Inteligencia Artificial es una comprensión preteórica de nuestras capacidades cognitivas. La habilidad para coordinar nuestra conducta con la de otros individuos descansa en gran medida en nuestra capacidad de adscribir deseos y creencias, pues suponemos que es el contenido de nuestros deseos y creencias lo que causa gran parte de nuestra conducta y, por tanto, lo que define qué comportamientos podemos esperar de nosotros mismos y de los demás. La fiabilidad de tales predicciones (y explicaciones) presupone que el comportamiento de los individuos responde a ciertas generalizaciones que relacionan, por un lado, deseos y creencias y, por otro, conductas. En el ámbito de la ciencia cognitiva, se denomina *psicología popular* a la teoría formada por ese conjunto de generalizaciones, si bien la fuerza explicativa de tales generalizaciones descansa en un compromiso ontológico de carácter mentalista, a saber:

La intuición mentalista (IM): Existen actitudes proposicionales (es decir, creencias, deseos, etc.) que tienen un contenido representacional, y que afectan y guían la conducta de quien las posee en virtud de ese mismo contenido.

En el seno de la Inteligencia Artificial se tiende a asumir IM, pues ¿quién se atrevería a dudar de su validez? ¿De qué otra intuición podríamos estar más seguros? Y, sin embargo, no deja de ser misterioso que la semántica de un estado mental pueda alterar la conducta de un organismo. Un ejemplo tomado de Fred Dretske [6] ilustra las razones para ser re-

6. Dretske (1988, 79-80).

ticentes. Imaginemos una soprano haciendo ejercicios de canto ante una copa, con el fin de romperla mediante el mero uso de la voz. Y así acontece al entonar nuestra soprano una de sus arias. Cuando el aria alcanzó la frase «No destruyas mi frágil corazón», la copa estalló. Coincidiremos en que el contenido del aria en poco ha afectado a la ruptura de la copa, sólo la altura de la voz parece contar. La cuestión que nos inquieta quizá ya se haya imaginado: ¿No ocurrirá lo mismo con los estados mentales que postula la psicología popular? ¿Cómo puede mover el mundo algo tan etéreo como el contenido, como la semántica? ¿No es, acaso, el contenido de nuestros estados mentales tan irrelevante para el movimiento de nuestro cuerpo como el contenido del aria para el estallido de la copa? ¿No tienen los contenidos mentales los mismos problemas que el alma cartesiana para afectar al (o dejarse afectar por el) cuerpo? El origen de esta sospecha deriva de una convicción que emerge con tanta fuerza como la intuición mentalista, a saber:

La convicción fisicalista (CF): El mundo es un sistema cerrado desde el punto de vista de sus propiedades físicas y, en consecuencia, cualquier hecho del mundo ha de tener una explicación física.

Si todos los fenómenos del mundo tienen una explicación física, cualquier explicación que demos de una conducta en términos mentales será, de algún modo, superflua, epifenoménica. Así, tomemos la siguiente generalización

(I) $M \rightarrow C$

donde «M» representa un estado mental o actitud proposicional, y «C» una determinada conducta. La convicción fisicalista obliga a que para cada generalización del tipo (I), haya otra generalización como la siguiente:

(II) $F \rightarrow C$

donde «F» constituye una propiedad física. Por tanto, debemos entender que es la propiedad física F y no la propiedad mental M lo que es causalmente responsable de la conducta C. Dicho de otro modo, aunque parezca que sea M lo que causa C, en realidad la causa de C es la propiedad física[7].

Sin embargo, el argumento que acabamos de presentar afecta no sólo a la relevancia causal de los contenidos mentales, sino a la eficacia causal de todas las propiedades de las ciencias especiales como la química, la biología, la geología, etc. En tal caso, los efectos de la convicción fisicalista serían devastadores y reducirían al epifenomenalismo a

7. Cf. Kim (1989, 1990, 1991).

todas las propiedades no físicas. Había que buscar algún remedio a tan nefasta situación. La solución parecía estribar en la posibilidad de encontrar algún tipo de dependencia entre F y M que fuese lo suficientemente fuerte como para respetar la convicción fisicalista, pero no redujese M a F, es decir, no redujese todas las propiedades a propiedades físicas, pues en tal caso la autonomía de las ciencias especiales quedaría aniquilada.

La idea de una teoría funcional contribuyó de manera crucial a la elaboración de esa noción de dependencia. En una teoría funcional los términos adquieren su significado por el rol causal que se le adscribe en la teoría a la entidad que designan. Las propiedades funcionales pueden instanciarse físicamente de varios modos. Objetos físicamente muy dispares pueden tener en común la propiedad funcional de «ser un freno» o «ser una llave de contacto». En consecuencia, la relación de realización es una relación unidireccional, pues no incluye la referencia a condiciones necesarias sino únicamente a condiciones suficientes. Diremos, por tanto, que una teoría funcional está físicamente realizada si pueden especificarse condiciones físicas suficientes de la realización de las propiedades que la componen. Una consecuencia de esta unidireccionalidad es que las propiedades funcionales no se reducen a las propiedades físicas y, en consecuencia, se puede reconocer la autonomía de las ciencias especiales. Sin embargo, la relación de dependencia debería ser lo suficientemente fuerte como para respetar la convicción fisicalista; pues nos proporciona, para cada propiedad funcional causalmente eficaz, un conjunto de propiedades físicas que constituye una condición suficiente de su instanciación. Ello garantiza que dos objetos o procesos con las mismas propiedades físicas compartan también sus propiedades funcionales y que todo fenómeno, aunque disponga de una explicación funcional, cuente también con una explicación física.

La otra cara de la noción de realización la constituye la noción de *sobrevenir*[8]. Así, podemos decir tanto que un conjunto de propiedades físicas de un objeto *realiza* una determinada propiedad funcional de ese objeto, como que tal propiedad funcional *sobreviene* a ese conjunto de propiedades físicas: sin alteraciones en la base física no puede haber cambios en las propiedades funcionales que sobrevienen a esa base física. Todo ello conduce, en el seno de la ciencia cognitiva, a una reformulación de la convicción fisicalista que resulta, en principio, compatible con la eficacia causal de las propiedades postuladas por las ciencias especiales. De modo más explícito, la nueva versión de la convicción fisicalista quedaría formulada como sigue:

8. Para una presentación de diferentes concepciones de la relación de sobrevenir, cf. Kim (1984, 1990).

El fisicalismo cognitivo (FCG): Una propiedad tiene relevancia causal si y sólo si:

(1) es una propiedad postulada por una teoría física o

(2) forma parte de una teoría funcional físicamente realizada, de tal manera que se puedan especificar las propiedades físicas que realizan (o a las que sobrevienen) las distintas propiedades que, desde esa teoría, se atribuyen al mundo [9].

Sólo las propiedades que satisfagan este principio podrán formar parte de una explicación científica; en consecuencia, la psicología científica únicamente podrá incorporar los compromisos ontológicos de la psicología popular si se puede probar que los estados mentales que esta última postula concuerdan con FCG. Precisamente, la relevancia filosófica de los distintos modelos de Inteligencia Artificial reside fundamentalmente en su habilidad para mostrar en qué medida FCG (heredero actual de CF) es compatible con IM [10]. De hecho, los defensores de la Imagen Sintáctica de la Mente entienden que el mejor argumento en favor de esa visión de la mente es que constituye el único modo conocido de conjugar FCG e IM [11].

III. LA IMAGEN SINTÁCTICA DE LA MENTE

Supongamos, pues, que los contenidos mentales se fijan en el seno de una determinada teoría funcional, a saber: la psicología popular. Hemos visto que, según el fisicalismo cognitivo, los estados que se postulan en el ámbito de una teoría funcional sólo son propiedades que alteran el mundo si se realizan físicamente. La noción de realización requiere que se indique cómo ciertas variaciones en las propiedades físicas de un organismo afectan a sus estados mentales. En este sentido, la Imagen Sintáctica de la Mente trata de mostrar cómo, para cada contenido mental M de un organismo O, hay un conjunto de propiedades físicas F de ese or-

9. Algunos textos básicos en la discusión actual acerca de las condiciones que debe satisfacer una propiedad para ser causalmente eficaz, son Fodor (1987, 1990), McLaughlin (1989) LePore y Loewer (1987, 1989), Horgan (1989), Kim (1989, 1990, 1991).

10. Esta es ciertamente una de las cuestiones centrales en ciencia cognitiva. Fodor (1987, 1985) es uno de los más fervientes defensores de la compatibilidad de IM y FCG, aunque Dretske (1981, 1987) y Millikan (1984) también se cuentan entre sus defensores. Dennett (1981, 1987c, 1987d) mantiene una posición más instrumentalista respecto a IM, mientras que P.M. Churchland (1986) y P.S. Churchland (1989) insisten en el conflicto entre IM y FCG. Para un mapa de las posiciones más relevantes respecto a esta cuestión: Fodor (1985), Dennett (1969, 1987a, 1987b, 1991) y también Lyons (1990a, 1990b).

11. Bajo el rótulo «Imagen Sintáctica de la Mente» se incluye la teoría representacional de la mente que apadrina Fodor. Se habla de Imagen Sintáctica para contraponer la teoría de Fodor a una teoría representacional derivada de los modelos conexionistas.

ganismo, tal que se cumple la siguiente relación de realización o (de sobrevenir):

$$F \rightarrow M.$$

Como vemos, este requisito exige que para cada distinción semántica haya una distinción física. Mas ¿cómo podría cumplirse esta condición?

Jerry Fodor considera que Turing y, con él, los modelos simbólicos realizan una aportación clave en este punto, pues muestran cómo *la semántica puede afectar al mundo a través de la sintaxis.* Es decir, el único modo de explicar la eficacia causal de los contenidos mentales es postular que éstos se hallan codificados en una estructura sintáctica. Esta es la Hipótesis del Lenguaje del Pensamiento propuesta por Fodor [12]. Ese lenguaje, al igual que los sistemas de lógica formal más sencillos, incluye un conjunto de criterios sintácticos para definir sus elementos primitivos, así como sus reglas de formación y derivación. Dos fórmulas de este lenguaje serán distintas en la medida en que difieran en su sintaxis. Hasta aquí nada nuevo. Lo que la Hipótesis de Lenguaje del Pensamiento añade es que:

> Toda distinción semántica en nuestros pensamientos encuentra su reflejo en la sintaxis del lenguaje del pensamiento, es decir, si dos pensamientos difieren en su contenido también difieren en su sintaxis. Ahora bien, si esto es así, ya tenemos todo lo que necesitábamos, pues no hay ningún problema en comprender cómo las distinciones sintácticas pueden intervenir en un proceso causal. Las distinciones sintácticas son distinciones formales y es fácil comprender, por ejemplo, cómo la geometría de una llave condiciona las cerraduras que con ella pueden abrirse.

Por tanto, el camino recorrido por la Imagen Sintáctica de la Mente, asociada a los modelos simbólicos, es el siguiente: Los contenidos mentales afectan al mundo porque

(1) Están codificados sintácticamente y
(2) tales distinciones sintácticas quedan realizadas en las propiedades físicas del cerebro.

Según nuestro esquema, tendríamos la siguiente relación de realización (o de sobrevenir):

Estados cerebrales — realizan → Fórmulas sintácticas
Fórmulas sintácticas = codifican ⇒ Contenidos mentales

12. Cf. Fodor (1975, 1987) y Fodor y MacLaughlin (1990).

con lo cual se habría mostrado cómo los contenidos mentales se realizan físicamente y, por tanto, cómo los estados de la psicología popular se realizan físicamente. De este modo, se puede responder ya a la pregunta acerca de cómo los contenidos mentales alteran el mundo, tienen relevancia causal: Los contenidos mentales alteran (y son alterados por) las propiedades físicas del mundo a través de la sintaxis. Para Fodor, la sintaxis es la nueva glándula pineal.

IV. DIFICULTADES PARA LA IMAGEN SINTÁCTICA DE LA MENTE

A pesar de su atractivo, la Imagen Sintáctica de la Mente no deja de plantear graves problemas. En primer lugar, la Imagen Sintáctica supone que hay principios para determinar el contenido mental que están asociados a cada fórmula sintáctica y, en definitiva, a los distintos estados cerebrales. Sin embargo, se ha argumentado poderosamente en contra de tal posibilidad, pues se considera que hay ciertos rasgos peculiares de los contenidos mentales (y, en general, de la semántica) que impiden tal correlación. En concreto, se suelen especificar dos rasgos: la normatividad y la relacionalidad. Este último hace referencia al hecho de que los contenidos mentales de un organismo no quedan fijados por su estructura física, sino por su relación efectiva con el entorno; hasta el punto de que dos organismos físicamente idénticos podrían tener contenidos mentales diferentes en función de las divergencias en sus respectivas biografías, en los entornos naturales y/o sociales en los que se han desarrollado. Por tanto, difícilmente puede haber un conjunto de propiedades cerebrales que sean condición suficiente de la instanciación de un contenido mental. Por otro lado, la normatividad hace referencia a la distinción entre correcto e incorrecto, a la posibilidad de cometer errores. Es esencial al pensamiento el poderse preguntar por si una inferencia es correcta o incorrecta, o si una creencia representa adecuada o inadecuadamente la realidad. Sin embargo, la noción de error tiene difícil cabida en el ámbito de estudio propio de la física, donde sería absurdo buscar principios normativos. El primer reto para la Imagen Sintáctica consiste, por tanto, en esquivar los escollos de la relacionalidad y la normatividad para mostrar que la secuencia cerebro → sintaxis ⇒ contenido mental realmente existe[13].

Con todo, estas no son las únicas dificultades que amenazan la viabilidad de la propuesta sintacticista. Hay indicios para defender que los

13. Para seguir la discusión en torno a la normatividad y relacionalidad de los contenidos mentales, cf. Burge (1979), Dretske (1988), Fodor (1987, 1990), McLaughlin (1989), LePore y Loewer (1987, 1989), Horgan (1989), Kim (1989, 1990, 1991), Pettit y McDowell (1986), Putnam (1975), Woodfield (1982).

modelos simbólicos sobre los que descansa la Imagen Sintáctica difícilmente pueden ser modelos que reflejen el modo en que efectivamente funciona nuestra mente. Los modelos sintácticos pueden simular muchas de nuestras capacidades cognitivas, pero no se ajustan al modo como opera nuestro cerebro. ¿Por qué?

Desde los modelos simbólicos, resulta sencillo explicar la productividad de nuestro pensamiento, es decir, la capacidad de generar un número indefinido de pensamientos a partir de una cantidad limitada de elementos. El carácter composicional de la sintaxis puede dar cuenta de este hecho. No obstante, parece que los modelos simbólicos fracasan a la hora de enfrentarse al problema del marco, es decir, a esa habilidad de los seres humanos (y de otros animales) para seleccionar, en un lapso muy breve de tiempo, los aspectos relevantes de una situación para el desarrollo de la actividad que se desea emprender. Cuando entramos en la cocina por la mañana para preparar el desayuno no necesitamos mucho tiempo para determinar que el color de los azulejos de las paredes no afecta al sabor del café con leche. Ante este problema, los modelos simbólicos tienen dos opciones igualmente insatisfactorias. O bien dotar al ordenador de un marco donde se preseleccionen los aspectos relevantes para cada situación o bien que, antes de cada actuación, el ordenador repase y tenga en cuenta todos los datos de que dispone. Esta última opción resulta inviable, pues impediría que el ordenador actuase en tiempo real, no podría simular la rapidez de nuestras actuaciones, al menos si respeta los límites biológicos en cuanto a la velocidad de las transmisiones cerebrales. Por otro lado, la opción de dotar al ordenador con un esquema nos obliga a pagar el precio de una excesiva rigidez que no concuerda con la flexibilidad de nuestras reacciones ante situaciones inesperadas, distintas de las tipificadas en el esquema. En resumen, los modelos clásicos para resolver el problema del marco o bien son demasiado lentos o bien son demasiado rígidos [14].

El problema del marco está vinculado a otras dificultades ante las que sucumben constantemente los modelos simbólicos como, por ejemplo, dar cuenta de la capacidad de generalización espontánea, del aprendizaje, de la degradación paulatina de nuestra actuación, etc. El desarrollo de los modelos conexionistas a principios de los 80 pareció arrojar alguna luz sobre estas cuestiones. A primera vista, se podía ver que propiedades tales como la generalización espontánea o la selección inmediata de los aspectos relevantes para un determinado curso de acción podían ser propiedades asociadas a modelos conexionistas [15].

14. Cf. Dreyfus (1979), Dennett (1983), Churchland y Churchland (1990).

15. Rumelhart y McLelland (1986, cáps. 1-4), Smolensky (1988), Clark (1989, cap. 6), Bechtel y Abrahamsen (1991, cap. 2).

V. EL SURGIMIENTO DE LOS MODELOS CONEXIONISTAS

Si los modelos simbólicos toman el lenguaje y su sintaxis como punto de partida para el análisis y simulación de la estructura del pensamiento, los modelos conexionistas buscan una mayor cercanía con la estructura del cerebro. Existe una gran variedad de modelos conexionistas, por lo que la descripción que sigue difícilmente podría hacer justicia a todos ellos [16]. Se trata más bien de destacar algunos rasgos que la mayoría de los modelos neurales comparten y que permiten entender su impacto sobre la Imagen Sintáctica de la Mente.

Los componentes fundamentales de los modelos conexionistas son las unidades y sus interconexiones. Las unidades pueden estar activadas o desactivadas, y pueden estar conectadas entre sí por nexos excitatorios o inhibitorios de distinto valor. El valor de esas conexiones y su carácter (excitatorio o inhibitorio) va cambiando a lo largo del proceso de entrenamiento de la red. El valor de activación de una unidad está en función de la fuerza global de los inputs negativos o positivos procedentes de otras unidades con las que se halla conectada, así como del sesgo que puede ir asociado a cada unidad. La influencia que una unidad A puede tener sobre otra unidad B es directamente proporcional, por tanto, al valor de activación de la unidad A junto al valor y naturaleza del vínculo que las une. Como es de esperar, las unidades están agrupadas en redes. En toda red mínimamente sofisticada cabe distinguir unidades de entrada, unidades ocultas y unidades de salida. Un modelo conexionista puede verse como una red de redes de unidades o pautas de conexión.

Un modo de comprender cómo estos sistemas pueden codificar y transformar representaciones es suponer que cada unidad representa un rasgo del mundo. Por poner un ejemplo muy simple, supongamos que hay tres unidades y cada una de ellas representa uno de los siguientes rasgos: «pelo», «vuela» y «pluma». Una conexión positiva entre «pluma» y «vuela» indica que la red cada vez que se activa «pluma» tenderá a activar «vuela», y viceversa. Una conexión negativa actuará, por el contrario, de manera inhibitoria: la activación de «pelo» inhibe «vuela». De este modo, la activación de una unidad opera como la formulación de una hipótesis acerca de la presencia de cierto rasgo en el mundo, mientras que el modo como están interconectadas las unidades revela en qué medida una hipótesis queda reforzada o amenazada por otras hipótesis.

De acuerdo con esto, no puede decirse que los contenidos proposicionales estén codificados en una determinada unidad de una red neural, sino en una red de unidades. Ciertamente, la misma red puede utilizarse para codificar una gran variedad de contenidos proposicionales. Las

16. Se puede encontrar una caracterización más detallada de los rasgos generales de los modelos conexionistas, así como una taxonomía de los mismos, en Rumelhart y McLelland (1986, cap. 2), Smolensky (1988), Hanson y Burr (1990), Bechtel y Abrahamsen (1991).

redes neurales están dotadas de una regla de aprendizaje que hace depender, por ejemplo, el tipo y valor de la conexión entre dos unidades de la frecuencia con que éstas se activan simultáneamente o, al menos, con un cierto grado de proximidad temporal. Esta regla permite que la red determine cómo están relacionados entre sí diferentes rasgos del mundo. Volviendo a nuestro ejemplo, si «pluma» y «vuela» tienden a activarse simultáneamente, la regla de aprendizaje dictará que se refuerce la conexión excitatoria entre ambas unidades, con lo que la red habrá aprendido que, si un objeto tiene plumas, es probable en un cierto grado que vuele; y algo semejante ocurre con las conexiones inhibitorias. De este modo, vemos que una red conexionista aprende modificando sus pautas de conexiones.

Una propiedad importante de las redes neurales es que no se paralizan si no pueden integrar coherentemente toda la información que reciben. Ocurre con frecuencia que los unidades inputs que se activan en un momento determinado representan rasgos del mundo que no son totalmente compatibles entre sí. Los modelos simbólicos se quedarían inmovilizados ante tal situación, a no ser que tal situación hubiese sido prevista con anterioridad y se hubiese programado una determinada solución. En cambio, los modelos conexionistas dibujan un paisaje de soluciones y seleccionan la que permite satisfacer mejor la información recibida, aunque la solución propuesta obligue a dejar de lado algunos de los datos presentados.

Los modelos conexionistas, por tanto, parecen estar en condiciones de dar cuenta de algunas de las capacidades cognitivas que más se resisten a un tratamiento simbólico. En primer lugar, hemos visto cómo puede dotarse a las redes neurales de reglas de aprendizaje que las lleva a generalizar espontáneamente cuando se detecta una cierta correlación (negativa o positiva) entre la presencia de varios rasgos del mundo. En segundo lugar, el comportamiento de las redes conexionistas, al igual que muchas de nuestras habilidades cognitivas, tiende a degradarse paulatinamente. Ello se debe a que el conjunto de condicionamientos que intervienen en el desarrollo de una tarea actúan de un modo «blando», es decir, cualesquiera de esos condicionamientos pueden violarse sin que la tarea se bloquee, si bien cuanto mayor sea el número y relevancia de los condicionamientos vulnerados mayor será el deterioro de la red en la realización de la tarea.

De este modo, se ha llegado a pensar que los modelos conexionistas podrían estar en condiciones de enfrentarse al problema del marco, sin tener que pagar ya sea el precio de la rigidez o de la lentitud. La dinámica de cancelaciones y activaciones que estas redes generan hacen más inteligible nuestra capacidad para enfrentarnos a las peculiaridades de cada situación. Las conexiones entre los distintos rasgos del mundo de los que tiene noticia el sistema están disponibles en las redes de conexiones, y el

sistema realiza un barrido de las mismas cada vez que desarrolla una tarea, con lo cual se seleccionan automáticamente los rasgos del mundo que pueden ser relevantes en una determinada situación. Este barrido puede realizarse en tiempo real porque, a diferencia de lo que ocurre en los sistemas clásicos, no se trata de ir trayendo al módulo central cada uno de los items de información archivados en la memoria, sino que cuando se activan determinadas unidades de salida, el impulso se propaga a lo largo de la red atendiendo a las conexiones que tales unidades de salida mantienen con el resto de las unidades del sistema, con lo cual puede recorrerse la totalidad del sistema en un número bastante reducido de pasos. De este modo, el sistema responde a los imperativos de cada situación sin atenerse a un esquema rígido, sino tomando en consideración toda la información sobre el mundo de que dispone; y, en segundo lugar, puede alcanzar una respuesta en tiempo real porque la dinámica de activaciones y cancelaciones permite recorrer la totalidad del sistema en un número bastante limitado de pasos.

VI. LA IMAGEN CONEXIONISTA DE LA MENTE Y EL PROBLEMA DE LA COMPOSICIONALIDAD

Parece, pues, que los modelos conexionistas están en mejores condiciones que los modelos simbólicos para enfrentarse al problema del marco, para dar cuenta de la capacidad de generalización espontánea, del aprendizaje, etc. Mas, si esto fuese realmente así, los modelos conexionistas estarían ofreciéndonos un análisis alternativo de cómo los contenidos mentales satisfacen FCG. A primera vista, sólo se hace necesaria una ligera modificación respecto a la solución propuesta por la Imagen Sintáctica de la Mente. En los modelos conexionistas los contenidos mentales no se codifican ya en fórmulas sintácticas, sino en redes de actividad. Por tanto, si nuestra mente fuese una red conexionista, entonces parece que el modo como se instancian los contenidos mentales no podría ser, como propone la Imagen Sintáctica, cerebro → sintaxis ⇒ semántica, sino

redes cerebrales — realizan → redes conexionistas
redes conexionistas = codifican ⇒ contenidos mentales.

Si esta solución fuese adecuada, la Imagen Conexionista de la Mente aparecería como una alternativa ante la Imagen Sintáctica, pues ambas surgirían como propuestas que pretenden hacer compatibles IM y FCG. En tal caso, la Imagen Sintáctica perdería gran parte de su atractivo, ya que el único argumento en su favor tenía la estructura de una inferencia a la mejor explicación. El surgimiento de una explicación alternativa so-

cavaría trivialmente la fuerza de tal inferencia. Además, la Imagen Conexionista no se presenta sólo como una explicación alternativa, sino como una explicación mejor y más fundamentada. Mejor, porque ofrece mejores perspectivas ante problemas como el del marco, ante los cuales los modelos simbólicos aparecen inermes; y más fundamentada, porque los modelos conexionistas no descansan exclusivamente en una inferencia a la mejor explicación, sino que responden a los rasgos fun damentales de la estructura del cerebro. Por tanto, si este diagnóstico se confirmase, la Imagen Conexionista vendría a sustituir a la Imagen Sintáctica.

Mas no está del todo claro que el diagnóstico vaya a confirmarse. De hecho, se insiste en que los modelos conexionistas fallan en un punto crucial: son incapaces de dar cuenta de la composicionalidad, sistematicidad y productividad de nuestro pensamiento y ello, como veremos, afecta también a la capacidad de la Imagen Conexionista de dar cuenta de la eficacia causal de los contenidos mentales. Desde este punto de vista, la Imagen Conexionista no sería ni siquiera una imagen alternativa y el papel que tendría reservado en el desarrollo de la ciencia cognitiva sería mucho más modesto.

Así, Fodor y Pylyshyn (1988) defienden que los modelos conexionistas no son incompatibles con los modelos simbólicos, pues se ubican en niveles explicativos diferentes. Mientras los modelos simbólicos analizan la estructura de la mente desde un punto de vista cognitivo, los estados que postulan los modelos conexionistas no son estados cognitivos, sino que, en todo caso, podrían contar como una posible implementación de los estados descritos por los modelos simbólicos. El argumento de Fodor y Pylyshyn depende de su concepto de «estado cognitivo»; para ellos, el caso prototípico de estado cognitivo es el tipo de estado que postula la psicología popular. Hasta tal punto que, según Fodor y Pylyshyn, un estado postulado por una determinada teoría es un estado cognitivo si y sólo si (1) se corresponde con un estado de la psicología popular y (2) tiene la misma estructura que los estados de la psicología popular. Lo que impide que los estados conexionistas sean estados cognitivos es este último punto. Fodor y Phylyshyn admiten que se pueda establecer una correlación sistemática entre estados conexionistas y estados psicológicos o representacionales, lo que dudan es que estos dos tipos de estados tengan la misma estructura. El razonamiento es bastante sencillo. Los estados de la psicología popular son productivos, sistemáticos y composicionales. Así, por ejemplo, somos capaces de generar (o de comprender) un número indefinido de pensamientos a partir de un número limitado de elementos primitivos. La única manera de explicar la productividad y sistematicidad del pensamiento es asumir que los mismos tienen *partes constituyentes*, de tal forma que el contenido de un pensamiento complejo se determine a partir del contenido de sus partes constituyentes.

La cuestión es, sin embargo, que los estados conexionistas no tienen partes constituyentes y el significado representacional de una red neural no puede determinarse a partir del contenido de sus partes constituyentes. Por tanto, los estados conexionistas no tienen carácter cognitivo, pues no se ajustan a la segunda de las condiciones definitorias de lo que es un estado cognitivo: en consecuencia, los modelos conexionistas no pueden dar cuenta, desde un punto de vista cognitivo, de la productividad, sistematicidad y composicionalidad del pensamiento. Se sigue, pues, que la Imagen Conexionista no es ni siquiera una imagen alternativa de la mente, dado que es incapaz de explicar los rasgos fundamentales de nuestra estructura cognitiva. Los modelos conexionistas, concluyen Fodor y Pylyshyn, ofrecen como mucho un modo alternativo de implementar los modelos simbólicos.

La respuesta no se hizo esperar. Los defensores de la Imagen Conexionista intentaron mostrar que la conclusión de Fodor y Pylyshyn no estaba fundamentada y, para ello, empezaron a desarrollarse modelos neurales cuyos estados constasen de partes constituyentes y, de ese modo, pudiesen dar cuenta de la productividad, sistematicidad y composicionalidad del pensamiento[17]. Smolensky (1987, 1991) propuso un modelo basado en el cálculo vectorial que permitía descomponer cualquier estado conexionista complejo en sus partes constituyentes. Se trataba de modelos conexionistas en los que estados como «taza con café» se representan de manera distribuida en función de la activación o desactivación de unidades más simples tales como «líquido caliente», «superficie curvada de porcelana», «asa para dedo», etc. Una red de este tipo podría representar estados como «café», «taza», «taza con café» y «taza sin café», según el estado de activación de las distintas unidades. En consecuencia, la representación conexionista de un estado como «taza con café» podría adoptar la forma de un vector <a.1, a.2, a.n> en el que constase el estado de activación de cada una de las unidades implicadas. Los componentes de ese vector formarían las partes constituyentes del estado «taza con café». Sin embargo, el modelo ha de complementarse con otro vector donde se definan las distintos roles estructurales que el estado «taza con café» puede jugar en las distintas representaciones. Las partes constituyentes del estado «taza con café» vendrían definidas por una operación del cálculo vectorial con los dos vectores mencionados anteriormente.

Disponemos, así, de un mecanismo que nos permite pasar de un vector complejo a sus partes constituyentes y viceversa. Según Smolensky, es posible determinar las especificaciones de una red neural que realizan un vector complejo, pero no se pueden fijar las especificaciones

17. Algunos de los principales modelos conexionistas son Pollack (1988), Hinton (1988), Smolensky (1987, 1991).

a las que sobrevienen cada una de sus partes constituyentes. Además, dado que hay múltiples maneras de descomponer un vector complejo, la determinación de las partes constituyentes de una representación será siempre relativa a un determinado proceso de descomposición. Smolensky no parece darle excesiva importancia a esta falta de unicidad en la descomposición, ya que estima que ello no conduce a un grado preocupante de indeterminación. No todas las estrategias descompositivas aportan los mismos resultados; algunas resultan más fructíferas que otras a la hora de predecir el comportamiento del sistema. Smolensky nos incita, por tanto, a dejarnos guiar por criterios operacionalistas a la hora de elegir el modo en el que los vectores complejos deban descomponerse.

Fodor y MacLaughlin (1990) manifestaron rápidamente su descontento ante la propuesta de Smolensky. Aun suponiendo que su propuesta de análisis vectorial fuese satisfactoria y que pudiesen descomponerse los estados conexionistas complejos en sus partes constituyentes, ello seguiría siendo insuficiente para dar cuenta de la productividad, sistematicidad y composicionalidad del pensamiento. La razón es que la propuesta de Smolensky es puramente instrumental, ya que, a pesar de ofrecer una manera de descomponer los estados conexionistas complejos en sus partes constituyentes, se ve obligado a reconocer que tales partes constituyentes no juegan ningún papel causal en la dinámica del sistema, pues no se pueden determinar las especificaciones de la red neural que realizan las partes constituyentes de un vector complejo. El análisis de Smolensky incrementa nuestra capacidad de explicar y predecir el comportamiento del sistema, pero no da cuenta del proceder real del mismo.

En consecuencia, el precio que Smolensky tiene que pagar para dar cuenta de la productividad, sistematicidad y composicionalidad del pensamiento parece excesivamente alto. El análisis de Smolensky no permite dar cuenta de cómo los estados mentales pueden satisfacer FCG, pues, si bien se puede mostrar cómo se realizan los estados conexionistas complejos, la propuesta de Smolensky no proporciona ningún criterio para determinar cómo se realizan las partes constituyentes de esos estados complejos. Hemos visto, sin embargo, que una parte importante de las transiciones causales entre estados mentales (como las vinculadas a la productividad y sistematicidad del pensamiento) sólo se puede explicar en virtud de la eficacia causal de sus partes constituyentes. Mas, según FCG, para que una parte constituyente de un estado mental sea causalmente eficaz (dé lugar, por ejemplo, a otro estado mental) es necesario que esté físicamente realizada. Por tanto, en la medida en que la propuesta de Smolensky es incapaz de mostrar cómo se realizan físicamente las partes constituyentes de un estado mental, está reconociendo que su planteamiento no puede mostrar cómo los estados mentales puedan ser causalmente eficaces y, en definitiva, cómo IM casa con FCG. De este

modo, si esta insuficiencia del planteamiento de Smolensky fuese expresión de una característica intrínseca del conexionismo, deberíamos concluir, con Fodor y MacLaughlin, que la Imagen Conexionista no puede sustituir a la Imagen Sintáctica, pues esta última sigue siendo el único modo conocido de dar cuenta de la eficacia causal de los contenidos mentales.

El debate no concluye aquí. Abogados de la causa conexionista, como Tim van Gelder (1990) y Andy Clark (1991), tienden a detectar una confusión crucial en la apelación que se hace a la composicionalidad del pensamiento para defender la Imagen Sintáctica de la Mente. Así, Van Gelder concede que hay un requisito genérico de *composicionalidad funcional* que todo tratamiento de la productividad y sistematicidad del pensamiento debe satisfacer. La composicionalidad funcional exige que exista un mecanismo fiable tanto para producir expresiones complejas a partir de número dado de elementos constituyentes como para descomponer las expresiones complejas en sus partes constituyentes. La satisfacción de este requisito de composicionalidad funcional puede garantizarse de varios modos. El más obvio es el que subyace a la Imagen Sintáctica de la Mente, a saber: una composicionalidad concatenativa. Este tipo de composicionalidad es el que permite reconocer en la fórmula compleja «((p → q) y p) → q)» los constituyentes «p», «q», « → », e «y». Es decir, utilizamos los mismos criterios para identificar estas partes constituyentes cuando forman parte de la fórmula compleja «((p → q) y p) → q)» que cuando se instancian aisladamente o formando parte de otra fórmula compleja. Tenemos un caso de composicionalidad concatenativa cuando las partes constituyentes se pueden reconocer en una parte de la fórmula compleja.

Todo lo que el argumento de Fodor y MacLaughlin muestra es que el análisis vectorial de Smolensky no satisface los requisitos de la composicionalidad concatenativa, pues en el estado conexionista que realiza una fórmula compleja no se pueden reconocer especificaciones que realicen las partes constituyentes de la fórmula en cuestión. Sin embargo, Van Gelder insiste en la posibilidad de diseñar modos de composicionalidad no-concatenativa, es decir, modos de composicionalidad que satisfagan los requisitos de la composicionalidad funcional sin que se haga necesario que se reconozcan las partes constituyentes en la fórmula compleja. Van Gelder propone como ilustración de composicionalidad no-concatenativa la posibilidad de asociar a cada fórmula de la lógica de primer orden un número de Gödel. Ello nos proporciona una función que nos permite determinar el número de Gödel de cualquier fórmula compleja a partir del número de Gödel de sus partes constituyentes, y a la inversa, según nos pide el requisito de composicionalidad funcional. Y, sin embargo, no se pueden reconocer en el número de Gödel de una fórmula compleja los números de Gödel correspondientes a cada una de sus par-

tes constituyentes, con lo cual tendríamos un modo de composicionalidad no-concatenativa.

Si aplicamos este caso a la propuesta de Smolensky, podemos ver que también su análisis vectorial podría satisfacer, en principio, las exigencias de la composicionalidad funcional de un modo no-concatenativo. En efecto, aunque no se pudiese reconocer la realización física de las partes constituyentes de una fórmula compleja en las especificaciones de la red neural que realiza esta última fórmula, podría definirse una función que nos permitiese pasar de la realización de la fórmula compleja a las especificaciones de las redes neurales que realizan sus partes constituyentes, y viceversa. De este modo, resultaría inteligible que las partes constituyentes de un vector complejo cumpliesen un papel causal, pues habría una función que, respondiendo a la dinámica efectiva del sistema, nos llevaría de la realización de las partes constituyentes a la realización del vector complejo, sin que haya ninguna necesidad de que en esta última realización quede un rastro reconocible de los estados conexionistas que son responsables de la realización física de las partes constituyentes.

De este modo, podemos también mostrar cómo el hecho de que los vectores complejos admitan varias descomposiciones no debería resultar en principio problemático. Como Smolensky mismo indica, no todas las descomposiciones son igualmente explicativas, algunas nos permiten predecir con mayor exactitud el comportamiento del sistema que otras. Pues bien, ahora cabría decir que, aunque los vectores complejos puedan descomponerse de varios modos, será una de esas descomposiciones la que responda a la dinámica efectiva del sistema, dado que sólo una de ellas dará cuenta de las transiciones efectivas entre los estados conexionistas que realizan las partes constituyentes y los estados que realizan el vector complejo. Lo que hizo que Smolensky se inclinase por una lectura instrumental de su análisis vectorial fue el confundir el requisito de composicionalidad funcional con las exigencias de la composicionalidad concatenativa.

Si esto fuese así, podríamos concluir que la Imagen Conexionista también puede dar cuenta de la eficacia causal de los contenidos mentales, ya que es capaz de compaginar IM y FCG. En consecuencia, la Imagen Conexionista aparecería como una imagen alternativa y la Imagen Sintáctica vería aniquilado el mejor de los argumentos en su favor, a saber: la inexistencia de una explicación alternativa. La Imagen Conexionista dejaría, con todo, abierta la posibilidad de que nuestra mente respondiese a un modelo híbrido en el que, en un entramado básicamente conexionista, ciertos módulos operasen según criterios simbólicos. En cualquier caso, conviene recordar que estamos ante una cuestión abierta y que estas consideraciones tienen una marcada provisionalidad.

VII. TEORÍAS COGNITIVAS, HOLISMO Y LA IMAGEN CONEXIONISTA

El surgimiento de la Imagen Conexionista no sólo afecta al esfuerzo por armonizar IM y FCG, sino que su efecto también alcanza a otras cuestiones en el ámbito de la filosofía de la mente. En esta sección, nos limitaremos a considerar dos efectos secundarios de particular interés. Así, veremos, en primer lugar, que el surgimiento de la Imagen Conexionista conlleva una visión más abierta y pluridimensional del concepto de teoría cognitiva que la que la Imagen Sintáctica de la Mente se ve obligada a preconizar. En segundo lugar, intentaremos mostrar que el desarrollo de la Imagen Conexionista requiere una reconsideración del debate en torno al holismo de lo mental.

Las teorías cognitivas son teorías funcionales. El hecho de que las teorías funcionales puedan definirse a diferentes niveles de abstracción, de manera que cada propiedad funcional pueda plasmarse físicamente de múltiples modos, hizo pensar que las teorías cognitivas deberían despreocuparse del detalle de la realización física y atender exclusivamente a los procesos que se detectan en el plano cognitivo. Esta opción venía a avalar, por otra parte, los programas de investigación en Inteligencia Artificial cuyos modelos parecían muy alejados de nuestras estructuras cerebrales. De este modo, se insistía en que las diferencias físicas entre seres humanos y ordenadores electrónicos no tenían por qué afectar a la naturaleza de sus capacidades cognitivas. Una expresión de este desentenderse de la realización es la insistencia de Fodor y Pylyshyn (1988) en que, si los estados conexionistas no están definidos en un plano cognitivo, el desarrollo de los modelos conexionistas es irrelevante para la comprensión de nuestra estructura cognitiva.

Los modelos conexionistas nacen, por el contrario, de la intuición de que el modo de realización es también relevante, que una buena teoría cognitiva debería incluir diferentes niveles de descripción [18]. De hecho, lo que ocurre a los niveles más bajos de descripción condiciona lo que puede ser verdad en los niveles cognitivos de descripción. Como vimos, uno de los datos que tiende a desacreditar a la Imagen Sintáctica como uno representación ajustada de cómo funciona nuestra mente es que sólo puede resolver el problema del marco al precio de violar los límites biológicos en cuanto a la velocidad de transmisión de señales y al número de unidades disponibles. Por el contrario, hemos subrayado que una de las virtudes de los modelos conexionistas es que, a primera vista, podrían permitir un tratamiento del problema del marco en términos biológicamente plausibles. En otras palabras, los límites que se fijan en el plano biológico definen los límites a los que debe ajustarse toda teoría plausible de la estructura cognitiva. Ello hace pensar que una teoría de la

18. Cf. Smolensky y otros (1992), Corbí (1993), Clark (1989, 1990), Cussins (1990).

cognición ha de incluir no sólo considerandos cognitivos, sino también referencias a lo que acontece a niveles más bajos de descripción. Parece, pues, que, frente a lo que se da a entender desde la Imagen Sintáctica, las teorías cognitivas han de articular diversos niveles de descripción, y no limitarse a los niveles cognitivos.

Por otro lado, la discusión en torno a la composicionalidad nos ha llevado a descubrir no sólo que las teorías cognitivas han de incluir varios niveles de descripción, sino que la relación entre tales niveles puede ser mucho más compleja de lo que desde la Imagen Sintáctica se pudiera pensar. Así, la relación entre niveles de descripción no tiene por qué responder a la simplicidad de la composicionalidad concatenativa, de manera que se puedan reconocer en la realización física de un estado cognitivo complejo las especificaciones que realizan físicamente cada una de las partes constituyentes de ese estado cognitivo complejo. La función que nos lleva de la realización física de las partes a la realización física del todo será, en general, bastante más compleja. Por tanto, podemos concluir que el surgimiento de los modelos conexionistas altera doblemente la concepción de la teoría cognitiva implícita en la Imagen Sintáctica pues, por un lado, obliga a incorporar diferentes niveles de descripción y, por otro, subraya la complejidad de las relaciones entre los diferentes niveles de descripción.

En segundo lugar, parece claro que la Imagen Sintáctica de la Mente es incompatible con el holismo del significado [19]. Si los contenidos mentales se atuviesen a criterios holistas de significado, difícilmente podría establecerse una correlación estable entre fórmulas sintácticas y contenidos mentales. Los criterios de individuación de las fórmulas sintácticas son composicionales y atomistas: la estructura sintáctica de una fórmula compleja es función de sus constituyentes más simples, y la contribución de estos constituyentes a la estructura compleja permanece estable, independientemente de la estructura compleja en la que se inserte. En cambio, si en la identificación de los contenidos mentales nos atuviésemos a criterios holistas, entonces el significado asociado a una fórmula sintáctica variaría en función de las alteraciones en el resto de los contenidos mentales del sistema, con lo que no podría mantenerse una correlación estable entre las propiedades sintácticas de una fórmula y las discriminaciones semánticas del contenido mental que a la misma se pudiesen asociar. Esta incompatibilidad entre la Imagen Sintáctica y el holismo de lo mental incitó a una lectura atomista de la psicología popular, dejando de lado los aspectos de la misma que vienen a subrayar su comportamiento holista [20]. Por el contrario, la Imagen Conexionista puede permitirse el lujo de recoger estos elementos holistas marginados

19. Cf. Fodor (1987), Fodor y Lepore (1992).
20. En defensa del holismo mental, cf. Davidson (1980), Dennett (1981, 1987e), Block (1986).

por la Imagen Sintáctica, ya que la manera de codificar y transformar representaciones de las redes conexionistas no puede ser atomista y, por tanto, no hay ninguna razón de principio que excluya la posibilidad de correlacionar sistemáticamente sus estados con los estados mentales que postula una lectura holista de la psicología popular[21]. En este sentido, pensamos que la Imagen Conexionista hace posible la rehabilitación del holismo presente en la psicología popular y que la Imagen Sintáctica se ve forzada a negar.

VIII. RECAPITULACIÓN Y CONCLUSIONES

La Imagen Sintáctica de la Mente, asociada a los modelos simbólicos, se propone como la única estrategia de que disponemos para dar cuenta de cómo IM pueda ser compatible con FCG. El surgimiento de los modelos conexionistas parece haber dado lugar a una estrategia alternativa que, además, encuentra mayor apoyo en la estructura del cerebro y permite enfrentarse a problemas cognitivos ante los que los modelos simbólicos parecen fracasar.

Así las cosas, el debate filosófico en torno a la Imagen Conexionista se centra, como hemos visto, en si esta nueva imagen constituye efectivamente una alternativa, es decir, en si realmente es capaz de mostrar cómo los contenidos mentales satisfacen FCG. Se objeta por parte de Fodor, McLaughlin y Pylyshyn que los modelos conexionistas no pueden dar cuenta, desde un punto de vista cognitivo, de la productividad y sistematicidad del pensamiento porque las redes conexionistas carecen de partes constituyentes, es decir, no son composicionales. Hemos intentado mostrar, a la luz de los comentarios de Van Gelder y Clark, que es necesario distinguir entre composicionalidad concatenativa y no-concatenativa, de modo que, en principio, no habría demasiada dificultad en reconocer la composicionalidad no-concatenativa de las redes conexionistas. De este modo, la Imagen Conexionista seguiría apareciendo como una imagen alternativa, por lo que constituirían una seria amenaza para la Imagen Sintáctica, dado que, en principio, los modelos conexionistas no sólo podrían dar cuenta de la productividad y sistematicidad del pensamiento, sino también de otros rasgos del mismo (como el aprendizaje, la generalización espontánea) ante los que los modelos simbólicos parecen naufragar.

No querríamos acabar, sin embargo, sin indicar que algunas cuestiones centrales en torno al problema mente-cuerpo parecen afectar por igual a los modelos simbólicos y a los modelos conexionistas. Si la Imagen Sintáctica tiene problemas para mostrar cómo la normatividad y la

21. Para una discusión de este punto, cf. Ramsey y otros (1990), Corbí (1993).

relacionalidad mental son compatibles con FCG, lo mismo acontece con la Imagen Conexionista. Ello nos hace pensar que la raíz de tales problemas se encuentra en un lugar más profundo y que, tal vez, las exigencias de FCG vayan más allá de lo que nuestras intuiciones fisicalistas nos dictan, por lo que parecería conveniente revisar las condiciones metafísicas que se imponen para reconocer la eficacia causal de los contenidos mentales[22].

BIBLIOGRAFÍA

Bechtel, W. y Abrhamsen, A. (1991), *Connectionism and the Mind*, Basil Blackwell, Oxford.

Block, N. (1986), «Advertisement for a Semantics for Psychology», en P. French y otros (comp.), *Midwest Studies in Philosophy*, v. 10, University of Minnesota, Minneapolis, 615-78.

Boden, M. (comp.) (1990), *The Philosophy of Artificial Intelligence*, OUP, Oxford.

Burge, T. (1979), «Individualism and the Mental», en P.A. French y otros (comp.), *Midwest Studies in Philosophy*, vol. 4, University of Minnesota, Minneapolis, 73-121.

Chomsky, N. (1957), *Syntactic Structures*, Mouton, La Haya. V. e.: *Estructuras Sintácticas*, Siglo XXI, México, 1974.

Chomsky, N. (1959), «Review of Skinner's *Verbal Behavior*»: *Language*, *35*, 26-58.

Chomsky, N. (1968), *Language and Mind*, Harcourt, Brace & World, New York. V. e.: *El Lenguaje y el Entendimiento*, Seix y Barral, Barcelona, 1974.

Churchland, P. S. (1986), *Neurophilosophy: Toward a Unified Theory of Mind/Brain*, The MIT Press, Cambridge, Mass.

Churchland, P. M. (1989), *A neurocomputational Perspective: The Nature of Mind and the Structure of Science*, The MIT Press, Cambridge, Mass.

Churchland, P. S y Churchland, P.S. (1990), «¿Podría pensar una máquina?»: *Investigación y Ciencia*, marzo, 18-24

Clark, A. (1989), *Microcognition*, The MIT Press, Cambridge, Mass.

Clark, A. (1990), «Connectionism, Competence, and Explanation», en Boden, 1990, 281-308.

Clark, A. (1991), «Systematicity, Structured Representations and Cognitive Architecture: A Reply to Fodor and Pylyshyn», en T. Horgan y J. Tienson (comp.), *Connectionism and the Philosophy of Mind*, Kluwer Academic Publishers, Dordrecht, 198-218.

Corbí, J. (1993), «Classical and Connectionist Models: Levels of Description»: *Synthese*, *95*, 141-168.

Cussins, A. (1990), «The Connectionist Construction of Concepts» en Boden, 1990, 368-440.

22. Queremos dejar constancia de nuestro agradecimiento a Tobies Grimaltos, Carlos Moya, Vicente Sanfélix y Josefa Toribio, cuyos comentarios nos han ayudado a mejorar versiones anteriores de este capítulo.

Davidson, D. (1980), *Essays on Actions and Events*, Clarendon Press, Oxford.
Dennett, D. (1969), *Content and Consciousness*, Routledge and Kegan Paul, London.
Dennett, D. (1981), «Intentional Systems», en *Brainstorms*, The MIT Press, Cambridge, Mass., 3-22.
Dennett, D. (1983), «Styles of Mental Representation»: *Proceedings of the Aristotelian Society*, *83*, 213-26. V. e., *La actitud intencional*, Gedisa, Barcelona, 1991, cap. 6.
Dennett, D. (1984), «Cognitive wheels: the frame problem of AI», en C. Hookway, *Minds, Machines and Evolution*, Cambridge University Press., Cambridge.
Dennett, D. (1987a), «Three Kinds of Intentional Psychology», en Dennett, 1987e, 43-68.
Dennett, D. (1987b), «Reflections: Instrumentalism Reconsidered», en Dennett, 1987e, 69-81.
Dennett, D. (1987c), «True Believers», en Dennett, 1987e, 13-36.
Dennett, D. (1987d), «Reflections: Real Patterns, Deeper Facts, and Empty Questions», en Dennett, 1987e, 37-42.
Dennett, D. (1987e), *The Intentional Stance*, The MIT Press, Cambridge, Mass.
Dretske, F. (1981), *Knowledge and the Flow of Information*, Basil Blackwell, Oxford.
Dretske, F. (1988), *Explaining Behavior. Reasons in a World of Causes*, The MIT Press, Cambridge, Mass.
Dreyfus, H.L. (1979), *What Computers Can't Do: The Limits of Artificial Intelligence*, 2.ª edición, Harper & Row, New York.
Dreyfus, H. L. y Dreyfus S. E. (1988), «Making a Mind versus modelling the Brain», en M. Boden, 1990, 309-333.
Feldman, J. A. y Ballard, D. H. (1982), «Connectionist models and their properties»: *Cognitive Science*, *6*, 205-54.
Fodor, J. A. (1975), *The Language of Thought*, Crowell, New York. V. e.: *El lenguaje del pensamiento*, Alianza, Madrid, 1984.
Fodor, J. A. (1983), *Modularity of Mind*, The MIT Press, Cambridge, Mass. V. e.: *La modularidad de la mente*, Morata, Madrid, 1983.
Fodor, J. (1985), «Fodor's Guide to Mental Representation»: *Mind*, *94*, 79-100.
Fodor, J. A. (1987), *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*, The MIT Press, Cambridge, Mass.
Fodor, J. A. (1990), *A Theory of Content and Other Essays*, The MIT Press, Cambridge, Mass.
Fodor, J. y Pylyshyn, Z. W. (1988), «Connectionism and Cognitive Architecture: A Critical Analysis»: *Cognition*, *28*, 3-71.
Fodor, J. y McLaughlin, B. (1990), «Connectionism and the Problem of Systematicity; Why Smolensky's Solution Doesn't Work»: *Cognition*, *35*, 183-204.
Fodor, J. y Lepore, E. (1992), *Holism: A Shopper's Guide*, Basil Blackwell, Oxford.
Hanson, S. y Burr, D. J. (1990), «What Connectionist Models Learn: Learning and Representation in Connectionist Networks»: *Behavioral and Brain Sciences*, *13*, 471-489.

Hebb, D. O. (1949), *The Organization of Behavior*, John Wiley & Sons, New York.

Hinton, G. E. (1988), «Representing part-whole hierarchies in connectionist networks»: en *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*, Montreal, Quebec, 48-54.

Hinton, G. E. y Anderson, J. A. (comp.) (1981), *Parallel Models of Associative Memory*, Lawrence Erlbaum, Hillsdale, New Jersey.

Horgan, T. (1989), «Mental Quasation», en J. Tomberlin (comp.), *Philosophical Perspectives, 3. Philosophy of Mind and Action Theory*, Ridgeview Publishing Company, Atascadero (California), 47-76

Kim, J. (1984), «Concepts of Supervenience»: *Philosophy and Phenomenological Research*, *65*, 153-76.

Kim, J. (1989), «Mechanism, Purpose, and Explanatory Exclusion», en J. Tomberlin (comp.), *Philosophical Perspectives, 3. Philosophy of Mind and Action Theory*, Ridgeview Publishing Company, Atascadero (California), 77-108.

Kim, J. (1990a), «Supervenience as a Philosophical Concept»: *Metaphilosophy*, *21*, 1-27.

Kim, J. (1990b), «Explanatory Exclusion and the Problem of Mental Causation», en E. Villanueva (comp.), *Information, Semantics, and Epistemology*, Basil Blackwell, Oxford, 35-55.

Kim, J. (1991), «Dretske on How Reasons Explain Behavior», en McLaughlin (comp.), *Dretske and His Critics*, Basil Blackwell, Oxford, 52-72.

LePore, E. y Loewer, B. (1987), «Mind Matters»: *The Journal of Philosophy*, *84*, 630-642.

LePore, E. y Loewer, B. (1989), «More on Making Mind Matter»: *Philosophical Topics*, *17*, 175-191.

Lyons, W. (1990a), «Intentionality and Modern Philosophical Psychology, I: The Modern Reduction of Intentionality»: *Philosophical Psychology*, *3*, 247-269.

Lyons, W. (1990b), «Intentionality and Modern Philosophical Psychology, II: The Return to Representation»: *Philosophical Psychology*, *4*, 83-102.

McCulloch, W. S. y Pitts, W. (1943), «A logical calculus of the ideas immanent in nervous activity»: *Bulletin of Mathematical Biophysics*, *5*, 115-33.

McLaughlin, B. P. (1989), «Type Epiphenomenalism, Type Dualism, and the Causal Priority of the Physical», en J. Tomberlin (comp.), *Philosophical Perspectives, 3. Philosophy of Mind and Action Theory*, Ridgeview Publishing Company, Atascadero (California), 109-136.

Millikan, R. G. (1984), *Language, Thought, and Other Biological Categories*, The MIT Press, Cambridge, Mass.

Minsky, M. A. y Papert, S. (1969), *Perceptrons*, The MIT Press, Cambridge, Mass.

Neumann, J. von (1956), «Probabilistic logics and the synthesis of reliable organisms from unreliable components», en C. E: Shannon y J. McCarthy (comp.), *Automata Studies*, Princeton University Press, Princenton, New Jersey.

Pettit, P. y McDowell, J. (comp.) (1986), *Subject, Thought and Context*, OUP, Oxford.

Pollack, J. (1988), «Recursive auto-associative memory: Devising compositional

distributed representations», en *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*, Montreal, Quebec.

Putnam, H. (1975), «The Meaning of 'Meaning'», en H. Putnam, *Philosophical Papers, 2: Mind, Language and Reality*, Cambridge University Press, Cambridge, 215-271.

Pylyshyn, Z. W. (1984), *Computation and Cognition: Toward a Foundation for Cognitive Science*, The MIT Press, Cambridge, Mass. V. e.: *Computación y Cognición*, Debate, Madrid, 1988.

Ramsey, W., Stich, S. y Garon, J. (1990), «Connectionism, Eliminativism, and the Future of Folk Psychology», en J. Tomberlin (comp.), *Philosophical Perspectives*, 4, Ridgeview, Atascadero, California, 499-533.

Rosenblatt, F. (1959), «Two theorems of separability in the perceptron»: *Mechanisation of Thought Processes: Proceedings of a Symposium Held at the National Physical Laboratory*, November, 1958, vol. 1, HMSO, London, 421-56.

Rosenblatt, F. (1962), *The Principles of Neurodynamics*, Spartan, New York.

Rumelhart, D. E., McClelland, J. L. y el PDP Research Group (1986), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, The MIT Press, Cambridge, Mass. V. e., parcial: *Introducción al Procesamiento Distribuido en Paralelo*, Alianza, Madrid, 1992.

Searle, J. (1980), «Minds, Brains, and Programs»: *Behavioral and Brain Sciences*, vol. 3, n.º 3, 417-458.

Searle, J. (1984), *Minds, Brains and Science*, Harvard University Press, Cambridge, Mass. V. e.: *Mentes, cerebros y ciencia*, Cátedra, Madrid, 1985.

Searle, J. (1990), «¿Es la mente un programa informático?»: *Investigación y Ciencia*, marzo, 10-16

Selfridge, O. G. (1959), «Pandemonium: A paradigm for learning», *Symposium on the Mechanization of Thought Processes*, HMSO, London.

Smolensky, P. (1987), «The Constituent Structure of Mental States: A Replay to Fodor and Pylyshyn»: *Southern Journal of Philosophy*, *26*, 137-160.

Smolensky, P. (1988), «On the Proper Treatment of Connectionism»: *Behavioral and Brain Sciences*, *14*, 1-74.

Smolensky, P. (1991), «Connectionism, Constituency and the Language of Thought», en B. Loewer y G. Rey (comp.), *Meaning in Mind: Fodor and his Critics*, Basil Blackwell, Oxford.

Smolensky, P., Legendre, G. y Miyata Y. (1992), «Principles for an Integrated Connectionist/Symbolic Theory of Higher Cognition»: *Informe CU-CS-600-92*, Computer Science Department, University of Colorado, Boulder.

Van Gelder, T. (1990), «Compositionality: A Connectionist Variation on a Classical Theme»: *Cognitive Science*, *14*, 355-384.

Woodfield, A. (comp.) (1982), *Thought and Object. Essays on Intentionality*, Clarendon Press, Oxford.

TEORÍAS DEL CONTENIDO MENTAL

Juan José Acero

I. INTRODUCCIÓN: INEXISTENCIA INTENCIONAL Y LA NORMATIVIDAD DE LO MENTAL

La intencionalidad, escribió el filósofo Franz Brentano (1838-1917), es el rasgo definitorio de los fenómenos psíquicos o mentales. Tal y como Brentano la concebía, la intencionalidad de un estado mental —de creencias y deseos, de recuerdos, expectativas o estados perceptivos— es una propiedad suya consistente en «estar dirigido a» un objeto. La propiedad es harto notable porque no es absoluto necesario que ese objeto exista de veras: se puede tener la experiencia de ver que hay un palo doblado en el agua sin que el palo esté realmente doblado; se puede creer que el rosal ha florecido sin que lo haya hecho en realidad, y se puede desear haber logrado el primer premio de la lotería sin que ello haya sucedido. En estos casos, alguien se ha forjado una representación de algo como si fuese tal o cual, sin que de hecho ese algo sea así. Sin embargo, en los casos de representación natural lo representado no tiene esa peculiaridad: es imposible tener ciertas manchas en la piel, entre otros síntomas, sin sufrir sarampión; es imposible que haya humo sin fuego. Haciéndose eco de esta diferencia y dando nueva vida a una doctrina añeja, Brentano sostuvo que el rasgo definitorio de la vida mental es la *inexistencia intencional* de sus objetos (Brentano, 1955/1957).

Un siglo después de que Brentano propusiera esta doctrina, se la enuncia de otra forma. En primer lugar, se dice, los estados mentales tienen naturaleza relacional: son relaciones entre agentes o sujetos psicológicos y contenidos. Así, la creencia de Ernesto de que el rosal ha florecido es una relación entre Ernesto y el contenido (o *proposición*) de que el rosal ha florecido. Y el deseo del gato de Elisa de capturar al ratón es-

condido en el seto es una relación entre ese animal y el contenido (o la proposición) de que el felino atrapa al ratón escondido en el seto. Creencias y deseos, expectativas y recuerdos son, en otra terminología, *actitudes proposicionales*, actitudes para con proposiciones o contenidos.

En segundo lugar, los contenidos de las actitudes proposicionales son entidades peculiares. Así, del hecho de que alguien, *A*, sea el portador de un estado mental, *M*, con un contenido sólo se sigue la existencia de *A*. (Puedo querer comprarme un yate o saludar a un agente de tráfico amable, sin que exista ningún yate que quiera comprarme o ningún agente de tráfico amable a quien quiera saludar, respectivamente.) En otros casos, ni del hecho de que *A* tenga una determinada actitud *M* hacia un contenido ni del hecho de que *A* no guarde *M* con ese contenido se sigue en ningún caso la verdad, y tampoco la falsedad, del contenido. (Crea o no Ernesto que Raúl es un espía, ninguna de ambas actitudes implica que Raúl sea un espía ni tampoco que no lo sea.) Además, los contenidos mentales parecen ser muy sensibles al punto de vista o a la perspectiva en la que se sitúa el agente. Aunque Edipo quería casarse con Yocasta —cuenta la tragedia clásica—, él no quería casarse con su madre. Edipo guardaba actitudes psicológicas del todo incompatibles con las proposiciones de estar casado con Yocasta y de estar casado con su madre. Pero no es nada fácil ver cómo puede ser posible tal cosa, ya que Yocasta y la madre de Edipo son una y la misma persona [1].

Propiedades como las acabadas de exponer han hecho de los contenidos mentales y, por consiguiente, de la cuestión de la intencionalidad un lugar en el que se cruzan los caminos de diferentes disciplinas. Así, la inexistencia intencional no sólo pone de manifiesto un fenómeno de interés lógico —como el que ilustran los ejemplos del párrafo anterior—, sino también uno psicológico. Las creencias pueden ser falsas, los deseos insatisfacibles y las percepciones ilusorias; pero aunque sus contenidos ilustren el fenómeno de la inexistencia intencional, o puede que por hacer tal cosa, unas y otros son el material con el que se cuenta en la explicación psicológica. Ahora bien, la explicación y la predicción de la conducta y la búsqueda de principios generales que invocar en esa tarea son objetivos que están justificados en la medida en que los estados mentales de los sujetos psicológicos son estados dotados de contenido y en que las leyes generales que se invocan al explicar la conducta de un agente son válidas de éste *en virtud de los contenidos de sus estados mentales*. El recurso al contenido es, entonces, inevitable. Así, por ejemplo, la conducta de Ernesto de diseñar un yate se hace inteligible porque desea poseer un yate muy especial, porque cree que ese género de yate no se en-

1. El *locus classicus* sobre las peculiaridades de la inexistencia intencional es el cap. 7 de Chisholm (1957). Excelentes son también el capítulo inicial de Dennett (1969) y el séptimo ensayo de Fodor (1981).

cuentra en el mercado y porque ese deseo y esa creencia encajan en una psicológica como ésta: *Para todo S, todo p y toda A: Si el sujeto S quiere que p, cree que la forma de lograr que p es realizando la acción A y nada obstaculiza realizar A, S tenderá a realizar A.* Pero las variables de esta ley, especialmente *p* (y de otras leyes como ésta), tienen entre sus dominios de valores contenidos de estados mentales y, por tanto, asumen la existencia de entidades como ésas.

Los contenidos mentales ocupan un lugar central en la explicación de la conducta. Sin embargo, interesan también por otra propiedad que poseen: la de la *normatividad de lo mental*[2]. Las creencias pueden ser verdaderas, pero pueden ser, asimismo, falsas; las percepciones son casi siempre verídicas, pero ocasionalmente son ilusorias; los deseos son a menudo consistentes, pero a veces se desea algo que no se puede obtener (o incluso algo lógicamente imposible de conseguir). En todos estos casos, el estado mental adolece de algún género de incorrección o carencia. Este hecho es de la mayor importancia. En el ámbito de la naturaleza, sea física, química o biológica, nada de ello se encuentra. Todo es como es; todo acontece como acontece. Si las leyes psicológicas sujetan los contenidos a condiciones generales y los contenidos tienen condiciones de verdad (y condiciones de satisfacción), la empresa psicológica no parece ocuparse de objetos naturales. Esta circunstancia sitúa el debate sobre la naturaleza del contenido mental en una perspectiva más amplia, tanto metodológica como filosóficamente: la de si es o no posible entender los hechos relativos al contenido mental en continuidad con los hechos de la esfera biológica. Una actitud naturalista entiende el contenido mental, el significado de la mente, como algo sujeto a los mismos principios generales que regulan el significado (o la representación) natural —el de los indicios o síntomas naturales, como el humo y las manchas en la piel—. Un programa de *naturalización* del contenido mental se propone mostrar que hay un conjunto de condiciones físicas necesarias, y globalmente suficientes, tales que, si un agente se halla en un estado corporal sujeto a esas condiciones, ese estado corporal tiene un cierto contenido. Descubriendo esas condiciones, se demostraría que lo intencional es parte de lo natural. Ahora bien, no es obvio cómo hacer cuadrar semejante programa con la inexistencia intencional de lo mental, señalada por Brentano. La tensión que aquí emerge es la que hace del Problema del Error (o Problema de la Disyunción) —¿es posible dar cuenta en términos naturalistas del hecho de que una actitud proposicional tenga un contenido incorrecto?— una cuestión decisiva en una teoría del contenido mental.

Un segundo par de fuerzas opuestas crea una nueva tensión en este dominio. Para los filósofos que piensan que los procesos y estados men-

2. Este rasgo del contenido mental es tenido especialmente en cuenta en la excelente panorámica ofrecida en Haugeland (1990).

tales son esencialmente distintos de los fenómenos no mentales, la intencionalidad de lo mental es un fenómeno que resulta de propiedades intrínsecas de la mente y, por ello, metafísicamente independiente de las circunstancias externas en que se da. No niegan ellos que los pensamientos puedan venir causados por circunstancias externas, pero insisten en que su contenido no depende de esas circunstancias, sino de rasgos *intrínsecos* del cuerpo o la mente de esos agentes. El mismo haz de propiedades intrínsecas —propiedades que un sujeto posee por sí mismo y no por guardar una relación con algo más— situado en un entorno diferente, seguiría dando lugar al mismo contenido y produciendo la misma vida mental. Esta es la característica diferenciadora del *internalismo* [3]. El internalista niega que puedan aplicarse a la vida mental los principios del significado natural, que son causales. Pero no puede olvidar que el contenido mental tiene condiciones de verdad (y condiciones de satisfacción) ni que éstas vienen determinadas por algunas de sus propiedades extrínsecas, por relaciones que guarda la mente (o el cuerpo) con las circunstancias externas. Así, parece haber aspectos del contenido que avalan un punto de vista *externalista*, un punto de vista según el cual el contenido mental es esencialmente dependiente del contexto: para el *externalismo*, un cambio de entorno implica un cambio en el contenido.

En la intersección de estos pares de fuerzas —naturalismo frente a no naturalismo, internalismo frente a externalismo— ha tenido lugar la confrontación entre las diversas teorías del contenido que han estado o están en liza en los últimos tiempos.

II. ANTIRREALISMO INTENCIONAL

Aunque el naturalismo, como he indicado, ocupa un lugar central en la concepción de los estados y procesos mentales de la mayoría de los filósofos de la mente actuales, las discrepancias en cuanto a la forma de ser naturalistas están a la orden del día. A grandes rasgos, dos puntos de vista se enfrentan entre sí: el del realismo y el del antirrealismo intencional. Son *antirrealistas intencionales* quienes piensan que la psicología de las actitudes proposicionales, la Psicología Popular, con sus estados y procesos, y con las leyes que sujetan éstos a condiciones restrictivas en virtud de los contenidos que posean, constituye una teoría radicalmente inadecuada de la vida mental; una teoría que ni siquiera cabe defender con la excusa de que acabará reduciéndose a una ciencia del cerebro.

3. La concepción cartesiana del contenido constituye una variante del principio general del internalismo, la variante más famosa a efectos históricos. Lo propio de ella es considerar, primero, que el contenido de los estados mentales viene determinado por propiedades intrínsecas de una segunda sustancia, la *res cogitans*; y, segundo, que tenemos (o podemos tener) un acceso directo a esos contenidos, un acceso mucho más seguro, en cualquier caso, que el que tenemos a los hechos de nuestro entorno.

Esta es la posición del *materialismo eliminativo* (véase Smith Churchland, 1980; Churchland, 1979, 1981; Smith Churchland y Churchland, 1983). Son también antirrealistas los que, como Stephen Stich, entienden que la adscripción de contenido es tan vaga y sensible al contexto y tan relativa al observador, que no hay en ella materia objetiva suficiente para sustentar leyes psicológicas válidas para diferentes sujetos en virtud del contenido mental que se les adscriba respectivamente. Un ejemplo de sensibilidad al contexto, que se discute en Stich (1983), tiene que ver con lo que Stich denomina el entorno doxástico de una actitud proposicional. Si la Sra. T. afirma: «McKinley fue asesinado» en pleno uso de todas sus facultades mentales y lo afirma de nuevo, mucho más tarde, cuando sufre de una seria pérdida de memoria y cuando ignora quién fue McKinley y qué es morir asesinado, ¿cree *lo mismo* la Sra. T. en ambas ocasiones? Para Stich, los partidarios de la psicología de la actitud proposicional no pueden aspirar a capturar similaridades en episodios tan distintos como éstos, porque cambios tan drásticos en el sistema global de creencias llevan consigo cambios de contenido igualmente drásticos.

Si bien doctrinas sobre el contenido mental como éstas representan de forma característica el punto de vista antirrealista [4], no son las únicas. Una variedad más sutil del antirrealismo tiene en Daniel Dennett a un portavoz destacado (véase Dennett, 1969, 1979 y 1987). El argumento más elaborado que Dennett ha presentado a este respecto (1987, 291-300) se apoya en las dos siguientes premisas: [P_1] Un artefacto únicamente posee intencionalidad derivada. [P_2] Los seres humanos somos artefactos: «Nuestra propia intencionalidad es exactamente como la de un robot» (298). En cuanto a [P_2], no es que Dennett no admita diferencias entre hombres y artefactos mecánicos, sino que considera que esas diferencias apuntan en la dirección de los procedimientos de diseño y manufactura. Mientras las máquinas son productos de nuestra inteligencia, los seres humanos somos «pensados» y construidos por nuestros genes egoístas. Pero eso no supone una diferencia decisiva, pues —al igual que opina Stich (1983, cap. 8)— Dennett también piensa que el cerebro [humano] es primero y sobre todo una máquina sintáctica.

En lo que respecta a [P_1], Dennett entiende que los estados internos de un artefacto (como un termostato o una máquina de bebidas) únicamente poseen significado o contenido en el contexto formado por los intereses de quienes lo diseñan, construyen y usan. Pero lo mismo, entiende Dennett, puede decirse de la intencionalidad humana. A su juicio, la única razón que autoriza a hablar de las creencias, deseos, intenciones y otros estados mentales de los seres humanos es que la atribución de actitudes como éstas permite muy a menudo explicar y predecir nuestra

4. Este punto de vista es tratado en profundidad en el capítulo 8 («El Eliminativismo y el futuro de la Psicología Popular») del presente volumen.

conducta. En esa medida, los seres humanos, los miembros de otras especies animales e incluso un computador, somos todos *sistemas intencionales*, es decir, sistemas a los que se aplica de manera fructífera una estrategia que Dennett califica de *intencional*. He aquí lo propio de esta estrategia:

> [...] primero uno *decide* tratar como agente racional al objeto cuya conducta ha de predecirse; luego, uno imagina qué creencias habría de tener el agente, dados su lugar en el mundo y sus propósitos. A continuación, y a partir de las mismas consideraciones, uno imagina qué deseos habría de tener, y finalmente uno predice que ese agente racional actuará para conseguir sus objetivos a la luz de sus creencias. Algo de razonamiento práctico a partir del conjunto de creencias y deseos elegido conducirá en muchos casos, no en todos, a una decisión relativa a lo que el agente debería hacer; es eso lo que uno predecirá que el agente *hará* (1987, 17).

Dennett enseña sus cartas aquí: que un agente sea un sistema intencional no significa que sea un sistema intrínsecamente intencional. Puede haber hechos que inviten a, que hagan razonable o que no dejen más camino abierto que, recurrir a esta apelación a actitudes proposicionales. Sin embargo, lo relevante del caso es que la estrategia intencional se sustenta sobre una decisión: la decisión de tratar a algo como un sistema intencional. La intencionalidad, los estados mentales con contenido, sólo existen en el ojo del expectador. No hay intencionalidad que pudiéramos llamar intrínseca o primitiva; toda ella es derivada. A esto se responde que los principios responsables de la armonía entre los estados sintácticos (o físicos) de un estado de un cerebro animal y los de una máquina son muy diferentes. En el primer caso, los principios operativos serían los de la selección natural y los del aprendizaje; en el segundo, los del diseñador y el ingeniero. La réplica de Dennett a esta objeción es que esta diferencia es simplemente ilusoria.

Para Dennett, entonces, la intencionalidad de la mente es *derivada*. Pero no sólo eso. Está, así mismo, *empíricamente indeterminada*. Cabría pensar que cuando los sujetos psicológicos son hablantes de una lengua, sus proferencias lingüísticas constituyen una garantía indiscutible de la existencia de actitudes proposicionales con un contenido determinado. Pero, como los materialistas eliminativos o como Stich, Dennett rechaza que esos contenidos —las proposiciones— tengan condiciones de identidad definidas y que esas proferencias tengan una única traducción correcta. Es decir, no sólo acepta Dennett la tesis de la *indeterminación de la traducción* que Willard V. Quine ha venido defendiendo desde los años sesenta [5], sino que entiende que la adopción de la perspectiva in-

5. El lugar clásico donde Quine expone esa tesis es Quine (1960). Para más información, véase la monografía titulada «El significado: la tradición escéptica», del volumen sobre filosofía del lenguaje de esta Enciclopedia.

tencional permite extender el dominio de validez de esa tesis desde el ámbito de las disposiciones al comportamiento verbal al de las disposiciones al comportamiento «interno» (véase Dennett, 1987, 40). Es decir, además de poder confeccionar diversos manuales de traducción de una lengua que sean compatibles por igual con las disposiciones lingüísticas de los hablantes de esa lengua e incompatibles entre sí, cabe también asignar diferentes sistemas de deseos y creencias a un mismo sujeto psicológico, de forma que ese sujeto se nos aparezca a esa luz como un agente racional, aunque esos sistemas sean incompatibles los unos con los otros.

III. EL REALISMO INTENCIONAL: LA PERSPECTIVA INTERNALISTA

El segundo punto de vista sobre la Psicología de la Actitud Proposicional (Psicología Popular o Psicología del Deseo/Creencia) es el de los partidarios del Realismo Intencional. Éstos creen que hay estados mentales dotados de propiedades intencionales genuinas y que esos estados se hallan involucrados causalmente, en virtud de su contenido, en la génesis del comportamiento. Además, entre los realistas intencionales algunos aceptan la existencia de leyes psicológicas que reconocen explícitamente esas propiedades intencionales. Y, todavía más, hay quienes aceptan que la Psicología Popular contiene algunas (muchas, todas) de esas leyes y que «carecemos de razones para dudar de —en realidad que tenemos razones sustantivas para creer— que es posible tener una psicología científica que reivindique la explicación de sentido común del deseo y la creencia» (Fodor, 1987, 16).

Más que la legitimidad de la Psicología del Deseo/Creencia, lo que más frecuentemente distingue a unos realistas intencionales de otros es la cuestión de la naturaleza de las actitudes proposicionales. Deseos, creencias y demás se entienden en ocasiones como relaciones entre agentes psicológicos y expresiones —en particular, «oraciones»— de un sistema de representación en el que la mente realiza sus cómputos (un *lenguaje del pensamiento* o mentalés). Así entendida, la creencia de Ernesto de que Raúl es un espía, es una relación entre Ernesto y la expresión #RAÚL ES UN ESPÍA#: la relación que se da por hallarse esta oración en la «caja de las creencias» de Ernesto. Filósofos como Jerry Fodor, Ned Block, William Lycan, Hartry Field y, antes que éstos, Wilfrid Sellars, encajan aquí. Para otros, como Brian Loar o Robert Stalnaker, que rechazan la hipótesis del lenguaje del pensamiento, las actitudes proposicionales son disposiciones *globales* que desempeñan una función mediadora entre las percepciones de los agentes y sus acciones. Si bien estas diferencias son importantes [6], en ambas facciones se acepta por igual que los estados

6. Panorámicas muy útiles a este respecto se ofrecen en Fodor (1990a) y en Haugeland (1990).

mentales son estados con contenido. Por lo tanto, las discrepancias sobre la naturaleza y los mecanismos de generación del contenido mental tienden a ser independientes de cómo se conciba el estado mismo. Sin embargo, la consecuencia obvia de que exista más de una concepción al respecto es que hay quienes proponen teorías del contenido de *representaciones* mentales y hay quienes proponen teorías del contenido de *estados* mentales.

Una forma muy popular de entender el contenido (o la semántica) mental es por referencia al lugar que ocupa en la organización mental del agente y, por lo tanto, por el papel que desempeña en la explicación de su comportamiento. Siguiendo a Dretske (1981, 202 s.), hablaríamos aquí de una aproximación *consecuencialista* al contenido de los estados mentales, una aproximación orientada hacia las consecuencias que produce la posesión de estados con esos contenidos. Los estados se ven, entonces, desde la perspectiva del sujeto a quien se le atribuyen. Decir cómo es el estado en que se encuentra el agente en una situación y un momento determinados es especificar cómo ve ese agente esa situación y su lugar en ella; y, por lo tanto, el factor responsable de la conducta subsiguiente del agente. Lo que el internalismo añade a esto es que ese factor depende de propiedades intrínsecas del agente. Y si ese internalismo tiene, además, vocación naturalista, añade también que el contenido depende de propiedades físicas intrínsecas del agente. El influyente ensayo de Hilary Putnam (1975) ilustra esta idea con el caso hipotético de las dos Tierras. Supongamos que en algún lugar del universo hay un planeta casi idéntico a la Tierra, un planeta en el que cada uno de nosotros tiene allí su doble físico. Siendo esto así, hay que aceptar también que nuestro doble es una réplica psicológica y lingüística nuestra. Óscar, un terráqueo que vivió en 1750, y su doble, que vivió en el correspondiente 1750, no sólo han pasado por las mismas experiencias fenomenológicas, sino que ambos las expresan por medio de las mismas palabras. Por ejemplo, ambos han bebido del líquido que en esos planetas se llama «agua» para saciar su sed, ambos utilizan ese líquido para lavarse y cocinar. Ambos, finalmente, en un cierto momento de sus vidas, ante un lago de su propio planeta, piensan que ahí, delante de ellos, hay agua. Los dos Óscar ignoran, sin embargo, que el líquido llamado «agua» en la Tierra es H_2O, mientras que el líquido llamado «agua» en la Otra Tierra es XYZ (pues estamos en 1750). Esta diferencia en la naturaleza del entorno no es obstáculo para entender la similitud de la conducta de los dos óscar. Vivir en un lugar donde hay H_2O y vivir en un lugar donde hay XYZ son propiedades extrínsecas de $óscar_1$ y de $óscar_2$, propiedades opacas del todo a efectos psicológicos. Una cuestión que suscita este experimento mental es, entonces, la de qué género de contenido es ése —el significado que «está en la cabeza»— que comparten dos sujetos que son físicamente idénticos y que es responsable de su comportamiento.

El experimento de Putnam legitima la idea de un significado que «está en la cabeza», siempre que éste cumpla dos *desiderata*: [D_1] Ha de sobrevenir a las propiedades físicas, intrínsecas o no relacionales, del cuerpo del agente. [D_2] Ha de ser pertinente para la explicación psicológica. Según [D_1], a identidad de propiedades físicas ha de corresponder identidad de contenidos mentales; y a diferencia de contenidos mentales, diferencia de propiedades físicas. Según [D_2], la invocación del contenido restringido habría de permitir la explicación causal de la conducta del agente. Además, parece razonable exigir un tercer *desideratum*: [D_3] que el significado que «está en la cabeza» sea una especie de significado (o de contenido). Los estados mentales son representacionales, significan la forma en que esta o aquella situación es para el agente. La cuestión central es si hay algo que responda a los *desiderata* [D_1]-[D_3].

En 1975, Putnam se limitó a caracterizar el contenido mental que «está en la cabeza» de la siguiente forma: del hecho de que un agente *A* se encuentre en un estado mental *M* con un contenido —por ejemplo, que *A* crea (desee, tema, etc.) que σ—, no se se sigue más que la existencia del propio *A*. Putnam acuñó el término «estado psicológico en sentido restringido» para referirse a un estado así entendido. Por extensión, se ha venido en denominar *contenido restringido* a un contenido de estas características: al contenido que «está en la cabeza», a lo que resta una vez que se cortan los vínculos que conectan el estado con los objetos y situaciones del entorno (es decir, una vez que se «ponen entre paréntesis» relaciones semánticas como la referencia, la denotación o la satisfacción). El contenido restringido es el género de contenido característico del *solipsismo metodológico*, el legado de las concepciones cartesianas (racionalistas y empiristas) de la mente de los siglos XVII y XVIII.

Tomadas en conjunto, las condiciones [D_1]-[D_3] dibujan una concepción *internalista* de los estados mentales que responde a las presiones naturalistas. Los estados mentales con contenido restringido representarían objetos, situaciones o estados de cosas en virtud de propiedades intrínsecas del cuerpo del agente. El problema es ahora el de qué contenidos satisfacen estos *desiderata*. Tres opciones serán brevemente descritas aquí.

(I_1) *Contenidos restringidos como condiciones de realización* (Loar, 1987, 1988a y 1988b). La idea principal de esta línea de investigación es que el contenido restringido σ de un estado mental *M* de un agente *A* representa no un estado de cosas real al que *A* se halla de alguna forma vinculado, sino la forma en que *A* concibe ese estado de cosas; no un aspecto del mundo real, sino un aspecto del *mundo nocional* del agente. Es por ello que se satisface [D_2]. Ese contenido σ, aun careciendo de condiciones de verdad, es una especie de contenido porque está sujeto a condiciones de otro género, a saber: a lo que Loar denomina *condiciones de*

realización. Es decir, el contenido σ restringido determina un conjunto de mundos posibles: el conjunto de mundos en los que sería el caso que σ. (El conjunto de los mundos en los que la creencia de *A* sería verdadera si *M* fuese una creencia; el conjunto de los mundos en los que el deseo de *A* se vería satisfecho si *M* fuese un deseo, etc.) [7]. Las condiciones de realización son constitutivas de contenido en virtud de un principio característicamente internalista:

> [...] es difícil comprender cómo puede verse uno a sí mismo concebir cosas sin concebir también en algún sentido «sobre» qué versan los propios pensamientos. Y a ello se le llama con propiedad contenido por un Principio de Transparencia del Contenido: si, desde una perspectiva no confusa, algo parece ser contenido, entonces es una clase de contenido (Loar, 1988a, 108).

Resta, entonces, conectar el cumplimiento de los requisitos [D_2] y [D_3] con el de sobreveniencia a las propiedades físicas del cuerpo de *A*. Y es a este respecto que la noción de *rol conceptual* aparece en escena. Pues el momento decisivo del presente enfoque llega con la idea de que un estado mental *M* de un agente *A* adquiere determinadas condiciones de realización en virtud del rol conceptual que posea *M* en la psicología general *A*. Esto quiere decir que el contenido restringido de *M* lo determina el *rol* (o *papel*) *causal* que *M* ejerce en la psicología de *A*: es decir, en el sistema de interacciones causales posibles de *M* con otros estados mentales del agente *A* (el rol funcional de *M* es su rol causal descrito en términos más abstractos que los de las ciencias del cerebro.) El rol conceptual es importante porque, entendido de la forma que se ha apuntado, garantiza que dos sujetos físicamente idénticos (en los respectos oportunos) poseerán el mismo sistema de principios causales internos, como les sucede a los dos Óscar de Putnam. Y si comparten un mismo sistema de roles causales, serán psicológicamente idénticos (en esos respectos). El corolario de ello es que la psicología del agente *A* sobreviene a las propiedades causalmente eficaces de su cuerpo. También [D_2] se cumple.

(I_2) *Contenidos restringidos en el lenguaje del pensamiento* (Sellars, 1963; Lycan, 1984, 1987; Block, 1986, 1991; Devitt, 1989). Supóngase ahora que la concepción representacional de la mente es (aproximadamente) verdadera: que *A* se encuentra en un estado mental *M* con contenido si, y sólo si, (i) *A* mantiene una relación *M* con una determinada

7. El concepto de mundo nocional, de Dennett, se invoca también para perfilar esta noción de contenido. Véase Dennett (1987). Sin embargo, mencionar en este contexto el nombre de Dennett no debe hacernos pensar que este autor sea un realista intencional. En cualquier caso, tanto las condiciones de realización como los mundos nocionales son contenidos *contextualmente indeterminados*: Dos sujetos que piensen «hoy hace sol» tienen en la cabeza el mismo contenido restringido, aunque su pensamiento acontezca en días diferentes.

oración del mentalés; y (ii) significa. Es posible, entienden algunos autores, matar dos pájaros de un tiro y dar cuenta tanto de la naturaleza de los estados mentales mismos como de sus contenidos restringidos en términos de la noción de rol conceptual. Así, en lugar de concebir —como en la opción (I_1)— el rol funcional de un estado mental *M* como el lugar que ocupa en el sistema global de estados, el nuevo enfoque introduce, si se lo compara con el anterior, dos variaciones de interés. La primera es que atribuye rol conceptual, en sentido estricto, a las *representaciones* mentales y considera que las propiedades causalmente eficaces de estos objetos son sus propiedades sintácticas o formales. Este es el Principio de Formalidad de los procesos mentales (Fodor, 1981, ensayo 8). La segunda es que para el rol conceptual de una representación importa la totalidad de sus conexiones con otras representaciones, con la estimulación sensorial (específicamente con los estímulos proximales) y con las rutinas motoras que desencadenan la conducta subsiguiente. Pues bien, la totalidad de estos vínculos en procesos perceptivos, en procesos cognitivos como la inferencia, tanto deductiva como inductiva, y la deliberación y en el control de la conducta determinan tanto el tipo de estado mental *M* del hablante como el contenido restringido que posea. Lycan expone lúcidamente la idea en estos términos:

> Soy de los que piensan que creer (en una u otra ocasión) es tener una representación-de-que-. A su vez, esto consiste en hallarse en un estado interno que pertenece a dos clasificaciones: el estado es una *creencia* de que por el tipo del rol funcional que desempeña, el tipo de tarea que lleva a cabo dentro de la burocracia interna de su agente. (Lo que hace de un estado una creencia como algo opuesto a un deseo o una intención es que el estado es una forma de almacenaje y/o parte de un mapa que sirve a su agente de guía para la acción, principalmente facilitando predicciones como resultados.) El estado es una creencia *de que* [...] por tener una cierta estructura interna en virtud de la cual sostiene relaciones inferenciales con otros estados de creencia reales o posibles y ciertas relaciones causales y/o teleológicas con cosas del mundo (Lycan, 1986, 161).

Si el rol conceptual de una representación determinase su contenido, el *desideratum* [D_3] se satisfaría. Lo mismo sucedería con [D_1], ya que el rol conceptual es una propiedad individual del agente, una propiedad que sólo depende de cómo esté internamente organizado el sistema de representaciones mentales del agente. Si, como es razonable suponer, ese sistema se halla físicamente materializado en su cuerpo, el rol conceptual sobreviene a las propiedades físicas. El decaimiento físico, o un accidente, podría alterarlo, con lo que se desvanecerían muchas de las conexiones entre representaciones y, por lo tanto, se alteraría de manera sustancial ese rol (como en el ejemplo de la Sra. T, que se consideró más arriba). Finalmente, el sistema de relaciones entre representaciones resulta pertinente para la explicación (causal) de la conducta (con lo que tam-

bién se satisface [D_2]). Imaginemos que durante un paseo por el campo alguien grita: «¡El toro está a punto de embestir a Ernesto!». De hecho, yo soy Ernesto, pero sufro momentáneamente de amnesia y no recuerdo mi nombre. Oigo el grito pero, quizás por pensar que ese tal Ernesto se halla lejos de donde yo me encuentro, no me muevo de donde estoy. Mi actitud se explica por el diferente rol conceptual que, dado mi estado, poseen las oraciones de mi mentalés (a) #EL TORO ESTÁ A PUNTO DE EMBESTIR A ERNESTO# y (b) #EL TORO ESTÁ A PUNTO DE EMBESTIRME#. Esa diferencia se debería al hecho de que de (a), pero no de (b), se sigue #ERNESTO ESTÁ EN PELIGRO#. No se sigue de (b) porque en mi «caja de creencias (activas)» no se encuentra la oración #ERNESTO = YO#.

(I_3) *El rol conceptual y el concepto de carácter*. La principal crítica dirigida contra esta manera de entender el contenido restringido es que inevitablemente conduce a la doctrina del *holismo* del contenido mental. Ello es así porque el rol conceptual de una representación depende constitutivamente de la totalidad de sus lazos con las demás representaciones, con la estimulación sensorial y con la conducta subsiguiente. Para algunos, el holismo del contenido es un hecho que hay que admitir sin más: cualquier variación en las creencias, deseos o percepciones del agente modificará ese sistema de nexos, es decir, alterará el rol conceptual de esa representación (véase Block, 1986, 1991). Para otros (Fodor, 1987, 1991b; Fodor y LePore, 1992), refuta el enfoque. A juicio de los segundos, una concepción holista del contenido restringido imposibilita que se lo invoque en la explicación psicológica, pues, siendo el rol conceptual del todo sensible a las peculiaridades de cada individuo por separado, no existirán leyes psicológicas que cuantifiquen sobre contenidos entendidos de esa manera y que sean válidas de cualesquiera agentes. La posibilidad de que sólo algunos de los ligámenes de las representaciones sean constitutivos de contenido y de que, entonces, el rol conceptual no sea holista —como se propone en Loar (1987) y en Boghossian (1992)— ha sido rechazada (por Fodor y LePore) aduciendo que presupone la distinción analítico/sintética, es decir, la distinción entre verdades que lo son sólo por el significado y verdades que lo son, además, por los hechos extralingüísticos. Y ésta es una distinción que se supone que Quine desacreditó del todo tiempo atrás (en Quine, 1951). Sin embargo, también se ha señalado, en Boghossian (en prensa), la incongruencia de que un realista intencional acepte argumentos típicamente antirrealistas como los de Quine.

Como alternativa a la vía que, presuntamente, conduciría al holismo, Fodor ha elaborado una teoría del contenido restringido que ha estado en el centro de la discusión durante los últimos años. Esa teoría se inspira en una que David Kaplan elaboró (1989) para las expresiones deícticas (como «yo», «aquí» o «ahora») y demostrativas (como «esa mujer» o

«aquellos árboles») y que daba cumplida cuenta de la dependencia contextual de este género de expresiones. La idea central de la propuesta de Fodor es la dependencia contextual del contenido restringido.

Según Kaplan, lo propio de las expresiones deícticas y demostrativas es la posesión de dos tipos de valores semánticos y la forma en que éstos se articulan entre sí. De un lado, estas expresiones tienen habitualmente *contenido* (o referente). Cuál sea éste es algo que depende del contexto en que se las use —de quién las profiere, del lugar y momento de tiempo en que se profieren, etc.—: «yo» refiere a mí, me tiene a mí por contenido, si soy yo quien profiere la expresión; a usted, si es usted quien la emplea, y así sucesivamente. Con «aquellos árboles» me refiero a los árboles que se ven desde mi ventana, si es a ellos a los que señalo o a los que miro al decir esas palabras; y me refiero a los árboles del bosque de Sant Llorenç, si es a ellos a los que señalo. Pero, además de su contenido, deícticos y demostrativos poseen otro valor semántico, que Kaplan denominó *carácter* (y Perry, *rol*; véase Perry, 1979). El carácter es la dependencia misma que guarda el contenido de la expresión con respecto al contexto en que se la usa, una regla que asocia contextos de uso a contenidos. Así, el carácter de «yo» es la función que asigna a cada contexto de habla en que se profiera este término el sujeto que lo profiere; y el carácter de «aquellos árboles» es la función que asigna a cada contexto en que el hablante señale unos árboles precisamente los árboles indicados.

La teoría fodoriana del contenido restringido identifica éste con una función análoga a la función-carácter de Kaplan [8], la que asigna a cada contexto o situación en que el agente psicológico se encuentre un contenido (o unas condiciones de satisfacción). Así, por ejemplo, el carácter de la representación #AGUA# es una función que empareja la Tierra con un líquido con la composición H_2O; que empareja la Tierra Gemela con un líquido con la composición XYZ; que empareja un tercer planeta con un tercer líquido posiblemente diferente de los anteriores; y así sucesivamente. La idea básica aquí es que lo que hace idénticos a los dos óscar, lo que da cuenta de la similitud de su conducta y, por ello, de la identidad de sus poderes causales ha de ser lo que compartan *a través de* los respectivos contextos. Y eso que comparten es la *función* citada: lo que haría que, si óscar$_2$ viajara a la Tierra y se le enseñase una gran porción de H_2O, pensara: #AHÍ HAY AGUA#; y lo que haría que, si fuese Óscar$_1$ quien viajara a la Tierra Gemela y allí se le enseñase una gran porción de XYZ, pensara también: #AHÍ HAY AGUA!#. Lo que distingue a unos contextos de otros, es decir, la especial composición de eso llamado «agua» por igual, no es pertinente para clasificar los estados mentales de los agentes de una manera útil para la psicología.

8. Algunas diferencias entre el carácter de Kaplan y el contenido restringido de Fodor se exponen en LePore y Loewer (1987) y en Stalnaker (1989).

Esta teoría satisface los *desiderata* [D_1]-[D_3]. Satisface [D_3] porque el contenido restringido es una especie de contenido. No es como el significado de una oración, pero sí como el significado de un deíctico o demostrativo: una función de contextos a algo más. Se satisface también [D_1], pues dos individuos físicamente idénticos habrán de estar dotados exactamente de los mismos mecanismos físicos en los que se materializa una y la misma función. Y se cumple el requisito [D_2], finalmente, ya que el contenido restringido únicamente atiende a las constancias *inter*contextuales, a esas propiedades de los entornos de los agentes que se traducen en propiedades intrínsecas de sus estados mentales. Así, la teoría de Fodor hace abstracción de las peculiaridades *intra*contextuales o extrínsecas que mantienen los agentes con sus entornos.

IV. LA NORMATIVIDAD DEL CONTENIDO Y EL EXTERNALISMO

No sólo las diversas maneras de entender el contenido restringido, sino la posibilidad misma de éste, han sido objeto de una crítica constante en los últimos tiempos. Algunas de las objeciones presentadas afectan a cuestiones técnicas de las teorías propugnadas. Así, por ejemplo, contra la teoría de Fodor del contenido restringido —opción (I_3)— se ha objetado (en Schiffer, 1990) que las funciones de contextos a condiciones de satisfacción son un constructo que no desempeña ninguna función teórica interesante; y que para que tanto la psicología científica como la popular lleven a cabo sus cometidos únicamente se precisan los roles causales de las expresiones del mentalés (es decir, las capacidades causales que poseen estas expresiones en virtud de sus propiedades sintácticas) [9]. Sin embargo, en la mayor parte de los casos las objeciones subrayan las limitaciones de que adolecen las concepciones internalistas para dar cuenta de la normatividad del contenido. Al margen de cómo se entienda el sistema intermedio entre la percepción y la acción —si constituido por estados globales del agente o por reglas que son sensibles a las propiedades sintácticas de las representaciones— ¿qué autoriza a entender los mecanismos *causales* de ese sistema como *razones* de los agentes? ¿Qué hace de ellas causas determinantes de un comportamiento que resulta ser *correcto* (o *incorrecto*)? El espacio de maniobra de que dispone un internalista para responder a esta pregunta es reducido. Puede apelar a la sistematicidad con que se asigna un conjunto de deseos o creencias a un agente; o puede recurrir a la coherencia interna que manifiestan el mundo nocional o las condiciones de realización de las actitudes proposicionales de un agente al que se supone una competencia máxima [10]. Todo ello

9. La imposibilidad de definir satisfactoriamente esa función ha sido señalada en más de una ocasión; por ejemplo, en Schiffer (1990) y Stalnaker (1989) y (1990).

10. A este respecto, es paradigmática la exposición que se hace en Haugeland (1990).

—se ha aducido como objeción— no basta (véase Stalnaker, 1989 y 1991). Un mundo nocional construido desde la perspectiva del agente puede ser todo lo coherente que se quiera y diferir arbitrariamente del tipo de entorno en el que se conforman las disposiciones a la conducta de ese agente. Es imposible especificar cuáles son las actitudes proposicionales de nadie sin adoptar supuestos muy complejos sobre la naturaleza de su entorno y de los vínculos que guarda el agente con él. Y aceptar esto supone renunciar al internalismo.

Esta objeción lleva consigo una respuesta a la pregunta de por qué habrían de importar la referencia de nuestros conceptos y las condiciones de verdad de nuestros pensamientos [11]. La respuesta es que no puede soslayarse ni la referencia ni la verdad, porque no es posible especificar el punto de vista del agente, el rol conceptual de sus representaciones mentales, sin indicar cómo se hallan conectadas éstas al entorno y cuáles son las condiciones bajo las cuales acuerdan con él. El contenido restringido tendría, entonces, un estatuto derivado que le conferirían esas conexiones y condiciones. La misma pregunta ha recibido otras respuestas. En Putnam (1978) se aduce que la referencia y la verdad, aunque ociosas a efectos de la psicología de los agentes —y muy en particular de su conducta lingüística—, son piezas claves de una teoría del *éxito* de ésta: de una teoría que se ocupe de por qué los conceptos de los agentes tratan del mundo y de por qué muchas de sus creencias son verdaderas y muchos de sus deseos son satisfacibles. Obviamente, el éxito de creencias y deseos no es un rasgo intrínseco de unas y otros. Pero ello no significa que no pueda haber buenas razones para buscar cerca de él los factores determinantes de su contenido. En particular, una posibilidad que ha encontrado un eco notable —véase, más abajo, la opción E_4— es la que entiende que los estados mentales son, al menos en condiciones óptimas, indicadores fiables del mundo, ligados a él por regularidades causales. Bajo cualquiera de estas opciones, —pero no sólo ellas—, se dispone de un canon externo para evaluar el contenido de conceptos y pensamientos.

Objeciones como las implícitas en estas últimas propuestas son sólo episodios dentro de una tradición que, si bien se remonta hacia atrás no mucho más de dos décadas, está respaldada por algunos argumentos que han ganado el favor de muchos y que han hecho concebir la posibilidad de desarrollar teorías anti-internalistas del contenido. Esos argumentos, y las teorías a ellos asociados, se basan en la idea —característica del *externalismo*— de que el contenido de los estados psicológicos de los agentes no es independiente del estado del mundo que hay más allá de sus superficies corporales. El nexo entre ambos determina que el contenido de un concepto consista en el referente que éste posea; y que el contenido

11. McGinn (1982, 220-229) contiene una amplia discusión de este tema.

de un estado mental consista en sus condiciones de verdad (o, más en general, en sus condiciones de satisfacción). A partir de aquí, la normatividad del contenido se explicaría por algún rasgo de las relaciones que guardan los agentes con su entorno, tanto natural como social. Obviamente, las concepciones externalistas rechazan el principio de que el contenido mental sobrevenga a las propiedades físicas intrínsecas del agente —es decir, rechazan [D_1]—. No por ello renuncian a [D_2], o sea, al requisito de que el contenido ocupe un lugar central en la explicación de la conducta, aunque en algún caso —como en E_4— ello suponga introducir un concepcto de conducta distinto del usual[12].

(E_1) *Teorías histórico-causales*. En sus conferencias de Princeton (en 1970), Saul Kripke arguyó decisivamente contra la idea de que sean las creencias que asociamos a un nombre propio «*N*» lo que determina que hablemos de un individuo (una persona, una ciudad, etc.), *x*, mejor que de otro diferente, *y*, al utilizar el nombre «*N*». En lugar del estado mental del hablante, Kripke esbozó una teoría *histórico-causal* de la referencia de los nombres propios. De acuerdo con ella, me refiero al filósofo griego al usar el nombre propio «Aristóteles» porque soy un eslabón de una cadena causal de comunicación que (i) se inicia con algo parecido a una ceremonia (explícita o no) de imposición de un nombre a un niño por parte de alguien; y que (ii) se prolonga cada vez que un hablante profiere ese nombre ante un interlocutor con la intención de referirse al individuo al que se refería la persona de quien él tomó el nombre «Aristóteles» (véase Kripke, 1980; Devitt, 1981). La *corrección* en la referencia no depende, entonces, de lo que «está en la cabeza» del hablante, sino del hecho *externo* de si forma parte de una cadena causal de comunicación y de las instituciones y formas de vida que sostienen esas cadenas[13].

El caso putnamiano de las dos Tierras, discutido más arriba, apunta en una dirección parecida, ya que ayuda a hacer tangible un tipo de estado psicológico puramente solipsista que sobreviene a las propiedades intrínsecas de los agentes. Sin embargo, Putnam pretendía subrayar también las diferencias semánticas que resultan de las relaciones causales entre el cerebro del agente y su entorno físico. Aunque en los cerebros de óscar$_1$ y óscar$_2$ se active por igual la oración-tipo #¡AHÍ HAY AGUA!#, no

12. Ni que decir tiene que [*D3*] deja entonces de ser crucial. El significado lingüístico siempre se entendió como una propiedad extrínseca de las palabras (una propiedad que no dependía en exclusiva del diseño gráfico del signo o de sus propiedades fonéticas). El contenido mental pasa ahora a verse de forma análoga sin que ello lleve a cuestionar su naturaleza semántica.

13. Aunque diferentes de la opción kripkeana, merecen destacarse, así mismo, las propuestas de Donnellan (1974) y Evans (1973) y (1982). Si nos atenemos a las fechas, la seccion X de Kaplan (1969) es un paso crucial en el desarrollo de todos estos enfoques. Sobre la teoría causal de Kripke y el trasfondo en que surgió, véase la monografía «La referencia», del volumen de filosofía del lenguaje de esta Enciclopedia.

piensan lo mismo, puesto que las condiciones bajo las cuales serían verdaderas las respectivos ejemplares de esa misma representación-tipo. La de óscar$_1$ es verdadera si, y sólo si, delante suyo hay H_2O; la de óscar$_2$ si, y sólo si, delante suyo hay XYZ. Las condiciones de verdad y de referencia de sus representaciones mentales difieren.

El factor aducido por Putnam para explicar esta diferencia apunta a la naturaleza de los entornos de los agentes psicológicos y a las circunstancias del aprendizaje de palabras como «agua», «limón», «león», etc. (o de la adquisición de los correspondientes conceptos) como factores responsables. Muy sucintamente enunciada, su propuesta es que lo que hace que nos refiramos *ahora* a una porción de una sustancia así, o a un miembro de una especie tal, son dos cosas: (i) que hayamos adquirido su uso por haber estado en contacto directo con porciones representativas de la sustancia o con miembros paradigmáticos de la especie en cuestión; y (ii) que ahora nos refiramos a algo que posea la misma composición o naturaleza que aquello que señalamos en las ocasiones iniciales. Pero la mismidad de composición o de naturaleza de una especie natural es independiente de nuestros deseos y creencias. (No tiene por qué «estar en nuestras cabezas».) Hay, así pues, un sentido en el que óscar$_1$ y óscar$_2$ no se encuentran en el mismo tipo de estado mental; pues, es de presumir, uno y otro óscar adquirieron «agua» en contacto con muestras de sustancias distintas. Esa es, entonces, la norma que dicta la corrección (o incorrección) y la verdad (o falsedad) de sus conceptos y pensamientos.

(E_2) *Anti-individualismo: el contenido mental como contenido socialmente determinado*. La invocación de relaciones causales cerebro-mundo en la determinación del contenido de un estado mental no es la única forma de rechazar el internalismo. Se puede hacer otro tanto cerrando el paso al *individualismo*, al punto de vista de que los factores responsables del contenido mental son factores —físicos o no[14]— que se encuentran en la esfera de la psicología individual, y no en el entorno del agente. Contra el individualismo, Tyler Burge ha presentado diversas variaciones sobre un argumento que ha ejercido una notable influencia (véase Burge, 1979, 1981 y 1988). El argumento se basa en un caso hipotético descrito en dos pasos. En el primero, nos concierne una mujer que, por sufrir de artritis, conoce por experiencia propia tanto los síntomas externos de su enfermedad como su fenomenología y que sabe usar el término «artritis» (o que ha adquirido el concepto #ARTRITIS#). En una ocasión, sintiendo dolores en el muslo, acude al médico y le comunica su temor de que la artritis se le haya extendido a un muslo. El médico la tranquiliza: la artritis no está extendiéndose, pues no afecta a

14. La concepción cartesiana de la mente asumiría, por su dualismo, una forma no materialista, no naturalista, de individualismo.

zonas musculares. En un segundo paso, consideramos una situación contrafáctica. La protagonista es la misma mujer de antes, con idénticas historia médica y experiencias fenomenológicas e idéntico estado físico. También ella siente un día dolores en un muslo y también le comunica a su médico su temor de que la artritis se le haya extendido a una zona muscular. Ahora, sin embargo, el médico le da la razón y anuncia, quizás, una intensificación del tratamiento. La diferencia entre ambas situaciones reside en que, en la segunda, «artritis» se usa también por la comunidad médica, y en otros ámbitos del medio social, para designar dolores reumatoides. Eso hace que no sólo el significado y la referencia de «artritis», sino también los contenidos de creencias y deseos dependan del entorno social. Las propiedades intrínsecas de la protagonista no varían de la situación de partida a la contrafáctica, pero la misma afirmación («ella cree que la artritis se le ha extendido al muslo») o el mismo pensamiento en primera persona (#LA ARTRITIS SE ME HA EXTENDIDO AL MUSLO#) tienen condiciones de verdad diferentes. Por lo tanto, habiendo variado sólo el medio social, el contenido de las mismas afirmaciones y pensamientos se ha modificado así mismo.

Merece la pena subrayar que el problema de la normatividad del contenido recibe un tratamiento satisfactorio aquí, en la medida en que uno atienda a factores que, como escribe Burge, están «"fuera" de un [agente] a quien se ve como organismo físico aislado, como mecanismo causal o como asiento de la conciencia» (Burge, 1979, 79). Esos factores, las costumbres y convenciones propias de la comunidad a la que pertenece el agente, fijan las pautas que determinan la corrección o incorrección de sus proferencias verbales y de sus pensamientos y deseos. Pero recurrir a estos factores en la consideración del contenido es más que renunciar a una concepción internalista de la mente —que es a lo que obliga la concepción de la adquisición de conceptos de género natural, de Putnam— es renunciar a una concepción individualista del contenido.

Estas renuncias dibujan los trazos más gruesos de una concepción de la mente y de sus estados completamente distinta de las de orientación internalista. La mente y sus principios de organización y evolución no son vistos ahora como instancias autónomas del entorno natural y social del sujeto psicológico, sino como una parte de esos entornos. Y el contenido mental como algo parcialmente «constituido» por los vínculos que unen a ese sujeto con sus entornos naturales y sociales. Así entendido, el contenido no es restringido (o estrecho), sino «contenido amplio»: es decir, un contenido que no sobreviene a las propiedades físicas intrínsecas del sujeto.

V. CONTENIDO Y EFICACIA CAUSAL

Entre las objeciones dirigidas contra teorías del contenido amplio como las acabadas de bosquejar, dos sobresalen en particular. La primera insiste en que el contenido amplio es ocioso dentro de la explicación psicológica (o cuando menos redundante). La explicación psicológica es explicación causal, pero dejaría de serlo si citase como causas a los objetos, situaciones y estados de cosas del entorno. Sólo «localmente» pueden ser éstos pertinentes para la psicología, a saber: en la medida en que afecten a las capacidades causales de los agentes. Y ello únicamente sucede a través de series causales de eventos particulares que se canalizan por las superficies corporales de estos agentes. Sólo las propiedades causalmente eficaces de los estados son responsables de sus contenidos. Este argumento presupone que las propiedades que confieren su identidad a un estado mental son las propiedades que lo hacen causalmente eficaz, es decir, las propiedades del estado que le otorgan un lugar en la explicación causal de la conducta. Sin embargo, se ha objetado que una cosa son las propiedades que fijan la identidad de un estado mental y otra diferente sus propiedades causalmente eficaces (véase Burge, 1986, 1989; Stalnaker, 1989). Así, las propiedades que hacen que el corazón de un hombre sea causalmente eficaz en el bombeo de sangre no son las mismas que las que hacen de él un corazón. Aquéllas son locales; éstas vienen fijadas por las relaciones que guarda el órgano en el conjunto de un cuerpo animal [15].

De otra parte, no sólo se ha insistido, en Burge (1986) y (1989), en que la psicología cognitiva es externalista —por ejemplo, tras considerar la explicación que de los mecanismos de la visión dio David Marr—, sino en que hay generalizaciones psicológicas con carácter de ley que cuantifican sobre contenidos amplios (Van Gulick, 1989; Braun, 1991) y en la eficacia causal de este género de contenido (Jackson y Pettit, 1988; Burge, 1989; Stalnaker, 1990; Braun, 1991). A este respecto, una crítica repetidamente hecha es que sí es posible que propiedades extrínsecas diferentes produzcan efectos también diferentes. Así, óscar$_1$ y óscar$_2$, ambos esta vez en un mismo planeta, piden: «¡agua!». Un interlocutor, conocedor del respectivo referente de cada proferencia, trae agua H_2O a óscar$_1$ y XYZ a óscar$_2$. Este contraejemplo lleva a proponer que se mida la identidad de contenido restringido por el siguiente rasero: dos creencias (deseos, etc.) M_1 y M_2 tienen el mismo contenido restringido si, para cualquier contexto C, M_1 y M_2 tienen idénticas capacidades causales en C (Fodor, 1987, capítulo 2). Entonces, frente a lo dicho hasta el momento,

15. Incluso se ha subrayado que una cosa son las propiedades de un estado mental causalmente eficaces y otra diferente las propiedades que la explicación causal invoca. Véase Van Gulick (1989); Burge (1989).

óscar$_1$ y óscar$_2$ no tienen en la cabeza el mismo contenido restringido, ya que los poderes causales de sus respectivos estados son diferentes.

Para superar esta objeción, Fodor ha reelaborado (1991) su argumento metafísico en favor del contenido restringido. Como en Fodor (1987), sigue exigiendo que las diferencias de contenido restringido se traduzcan en diferencias de poder causal, pero Fodor añade ahora la condición de que «la diferencia en las causas no esté conceptualmente ligada a la consiguiente diferencia en los efectos» (1991, 21). Medidas por este rasero, las diferencias entre óscar$_1$ y óscar$_2$, cuando ambos piensan #¡AHÍ HAY AGUA!#, no son diferencias de contenido restringido. La petición de óscar$_1$ es una petición-de-H_2O, mientras que la de óscar$_2$ es una petición-de-XYZ; la conducta subsiguiente sería, por lo tanto, la de satisfacer una petición-de-H_2O, en un caso, y la de satisfacer una petición-de-XYZ, en el otro. Pero el vínculo entre las peticiones y las conductas subsiguientes sería aquí *conceptualmente necesario*. No es el vínculo que existe entre dos trozos de roca, uno de los cuales es un meteoro y el otro no, cuando golpean la superficie de un planeta. El primero produce un cráter profundo al impactar con ella; el segundo apenas si levanta un poco de polvo. No hay razón para dudar de la eficacia causal de ciertas propiedades extrínsecas (o relacionales), pero no todas están cortadas por el mismo patrón. El ejemplo de las dos rocas sugiere más bien —siguiendo a Burge (1989)—, que mejor que los efectos de ciertas propiedades, lo relevante podrían ser sus antecedentes causales; y que en la medida en que la prueba de Fodor clasifica los estados mentales por sus efectos en contextos potenciales, es insensible a las diversas formas en que las ciencias especiales reconocen poderes causales.

VI. EL REALISMO INTENCIONAL Y LA NATURALIZACIÓN DEL CONTENIDO: TELEOLOGÍA, INFORMACIÓN Y DEPENDENCIA ASIMÉTRICA

La segunda objeción a propuestas como las de Kripke y Burge es que son incompatibles con el proyecto de naturalización del contenido. Lo son, se aduce, en la medida en que la elucidación del contenido de un estado mental pone en juego nuevos recursos intencionales (como la intención de preservar el referente en cada eslabón de una cadena causal de comunicación); o en la medida en que apela a normas sociales. Hay autores, como Burge, que entienden que la búsqueda de teorías naturalistas del contenido obedece a prejuicios materialistas; otros, como Fodor, Ruth Millikan o Fred Dretske han optado por seguir esa búsqueda. He aquí, entonces, algunas opciones en esta dirección.

(E_3) *El contenido amplio desde una perspectiva teleológica*. Para las *concepciones teleológicas*, las funciones biológicas de los mecanismos que

generan representaciones son el elemento determinante en la fijación del contenido que éstas tengan. Es decir, son las funciones biológicas las que determinan las condiciones de identidad de los estados mentales. Las funciones biológicas de esos mecanismos se explican, según la ortodoxia, en términos de su historia evolutiva. Poniendo un ejemplo, existiría una íntima conexión entre el mecanismo que lleva a los castores a golpear su cola contra el agua, para avisar de la presencia de peligro, la historia evolutiva de ese mecanismo y el hecho de que esa conducta *signifique* la presencia de un peligro potencial. La selección natural sería la instancia invocada para explicar por qué ese mecanismo está presente en la conducta de la especie.

Las aproximaciones teleológicas reúnen dos méritos incuestionables. En primer lugar, la naturalidad con que permiten abordar el problema de la normatividad de lo mental. Tan pronto como uno explica el contenido de una representación a partir de la función que ésta realiza, se ha admitido la posibilidad de que esa función se lleve a cabo bien o mal; que las funciones se ejerzan propiamente o que haya disfunciones o malos funcionamientos. La correcta realización de la funciones es aquí el canon básico de normatividad.

En segundo lugar, las aproximaciones teleológicas son muy flexibles a la hora de hacer frente al Problema del Error a propósito del contenido. (Esa flexibilidad es precisamente considerada por algunos la limitación mayor del enfoque.) Este es el problema de decidir cuál de entre diversos contenidos alternativos posee una representación (o estado mental) de un organismo biológico. Las ranas, por ejemplo, reaccionan a la presencia de insectos en su hábitat natural proyectando hacia ellos sus lenguas, y capturando así su alimento. Pero la representación o el estado interno responsable de esa conducta puede significar, en principio, tanto #INSECTO# o #MOSCA# como #MANCHA NEGRA QUE SE MUEVE# o incluso #ALIMENTO#. Ya que cualquiera de estos significados puede corresponder por igual a la función desempeñada por el mecanismo responsable, el contenido del estado (o la representación) mental estaría indeterminado. Sin embargo, hay varias maneras de matizar el principio que comparten las concepciones teleológicas y poner coto a esta indeterminación. Se puede vincular la función biológica a las clases naturales presentes en el entorno, como han propuesto Sterelny y McGinn. La representación interna de la rana significa, entonces, #INSECTO#, #MOSCA#, #ABEJORRO#. Se puede sugerir, como lo ha hecho Neander, que los principios evolutivos seleccionan aquellos rasgos que son responsables de las propiedades fenoménicas que proyecta el entorno en el individuo. La representación significaría, entonces, algo diferente: algo como #MANCHA NEGRA QUE SE MUEVE#. O se puede hacer depender el contenido de la representación del empleo que haga su usuario para satisfacer sus necesidades, como en la compleja e inte-

resante propuesta de Millikan. En ese caso, la representación significaría #ALIMENTO# [16].

(E_4) *El contenido amplio desde una perspectiva informacional-causal.* La principal alternativa a las teorías teleológicas del contenido es la que proporcionan las *teorías causales de la indicación*. Su rasgo esencial consiste en entender los estados representacionales como estados de sistemas que *procesan información*: de sistemas que son capaces de detectar en su entorno la presencia de información, de recuperarla y de hacerle desempeñar un papel decisivo en el control de la propia conducta. Pero la detección, recuperación y transformación de la información, en la que reparan las teorías causales, son funciones biológicas, funciones para cuya ejecución la evolución natural ha seleccionado mecanismos en las especies. Los enfoques teleológico e informacional hacen una buena parte del camino juntos (van Gulick, 1980, es un ejemplo de esta forma de simbiosis). El punto a partir del cual ambos enfoques se separan se sitúa más allá de los mecanismos de representación innatos. En lo que concierne a éstos, los dictámenes de las concepciones teleológicas resultan razonables. Sin embargo, una objeción que comúnmente se les hace a estas últimas concepciones es que adscriben funciones biológicas a las creencias (y deseos) de los seres humanos —y en general todos los estados representacionales que resulten de procesos de *aprendizaje*—, cuando no es obvio ni que las tengan ni que estos estados posean ellos mismos historias evolutivas (véase Fodor, 1990, 65-69).

Las teorías informacionales, por su parte, tratan el contenido mental como un tipo de significado natural no absolutamente distinto del que encontramos en los índices naturales (el humo, señal de fuego; ciertas manchas, síntoma de sarampión). En todos estos casos, algo, R, se convierte en representación de otra cosa, σ y, por ello, en *portador de la información* de σ que cuando se establece un vínculo nomológico entre la aparición de R y la presencia de σ [17]. Pues bien, como ha escrito Drestke, un sistema dotado de mente es uno que procesa información en una estrecha y exitosa interacción con su entorno:

16. Una interesante panorámica de las concepciones teleológicas se ofrece en Agar (1993), que contiene, además, una solución alternativa. Véase también en el presente volumen la monografía titulada «La concepción teleológica de los estados mentales y de su contenido», que desarrolla en profundidad este género de aproximaciones. La crítica más detallada de estos enfoques se encuentra en el capítulo 3 de Fodor (1990).

17. El vínculo nomológico tiene rango de una ley causal, lo cual significa que ciertos enunciados contrafácticos (o probabilísticos) deben ser verdaderos: enunciados como «Si fuese el caso que se incendiara un bosque, se produciría R la presencia de humo»; o como «La probabilidad de que se produzca el evento R, dado el evento de que, es 1». (Véase Van Gulick 1980; Dretske 1981, 1988a, 1988b; Stampe 1979; Stalnaker 1984, 1989 y 1990).

> Pienso que uno tiene que hacer encajar un cuerpo *en* el mundo de una forma muy íntima para tener una mente. ¿Por qué? Porque hablar de la mente —en particular, hablar de lo que uno piensa, desea e intenta (*no* de los pensamientos, los deseos y las intenciones mismas)— es hablar de esa red de relaciones, principalmente relaciones *entre* el cuerpo y lo que le rodea, que ayuda a explicar la interacción exitosa del cuerpo con, y su integración en, eso que le rodea (Dretske, 1988b, 55).

La aplicación de este principio conductor al caso de la creencia aprendida se hace posible cuando se entiende el aprendizaje como un proceso a lo largo del cual un sujeto *A* se ve expuesto a estímulos y señales que llevan consigo informaciones muy diversas, σ*, σ**. A lo largo de ese proceso, se llega a reclutar una cierta estructura física (o estado interno), *R*, —una asamblea de neuronas de *A*—, una estructura sensible, que responde de forma selectiva, a la presencia de la información de que σ. En ese momento, el sujeto *A* ha alcanzado a disponer de representación *R* con el contenido σ (o ha adquirido la disposición a creer que σ en condiciones específicas). Lo que así emerge es una nueva ley causal: la ley de que (en condiciones óptimas) si fuese el caso de que σ, se activaría la estructura *R*; y de que (en condiciones así) si se activase la estructura *R*, sería el caso que σ. Frente a las concepciones consecuencialistas del contenido, como las ilustradas más arriba, este otro enfoque tiene un cariz claramente *etiológico* (Dretske, 1981, 201 s.): es el origen, y especialmente el origen informacional, de las estructuras (o los estados) representacionales lo que fija su contenido. Si una estructura *R* del cerebro se ha hecho sensible a la presencia de pirámides tras haber sido expuestos a esas construcciones —es decir, si se ha generado una ley causal «pirámide → R'»—, esta estructura puede identificarse con una representación mental de las pirámides (o con un concepto de pirámide).

Desde la presente perspectiva, el contenido no sobreviene a las propiedades intrínsecas del portador de la representación (o del estado) mental, sino a propiedades relacionales, etiológicas, de las estructuras semánticas mismas. Pese a ello, ese contenido no es un epifenómeno dentro de la vida mental del agente, sino un ingrediente (entre otros) responsable de su comportamiento. Esta última idea precisa, sin embargo, de una forma especial de concebir la conducta del agente. Si por episodio de la conducta intencional de un agente *A* se entiende un movimiento (o una serie de movimientos) *G* controlado por una estructura *R* (de *A*), entonces parece que el contenido (o significado) de *R* es tan sólo un epifenómeno de *G*, pues son las propiedades físicas de *R* lo que (junto a otros factores causales) produce *G*. Pero si por episodio de la conducta de *A* no se entiende una instancia de *G*, causada por *R*, sino *la causación de G por R* —como se hace en Dretske (1988a), (1988b)—, entonces que *R* tenga contenido ya no es un epifenómeno. Ahora, para dar cuenta de por

qué *R* causa *M*, de por qué *R* ha adquirido control sobre *M*, hay que apelar por fuerza a la etiología de *R*, a los eventos o estados de cosas externos representados en virtud de los vínculos nomológicos que ligan a *R* con éstos.

Pues bien, esta concepción ha sido acusada de comprar el contenido mental a un precio demasiado barato. Ello es así, se dice, por la *robustez* del contenido o significado mental: porque *R* puede ser activada por muchas otras cosas que carecen de las propiedades pertinentes. Dos son los casos que han de distinguirse. (i) Que #PIRÁMIDE# sea activada por cubos. Entonces #PIRÁMIDE# indica la presencia de cubos, y estaríamos ante un caso manifiesto de representación *errónea*: aunque fuesen cubos lo que tuviésemos delante, estaríamos identificándolos con pirámides (pues la ley causal «cubo → #PIRÁMIDE#» sería operativa). Y (ii) que, por disponer de información colateral, pensásemos en pirámides al ver una efigie de Amenofis IV. En este caso, el vínculo causal «efigie de Amenofis IV → #PIRÁMIDE#» sería responsable de esa derivación de nuestro pensamiento, porque la efigie indicaría la presencia de pirámides. Posibilidades como éstas, se aduce, son obstáculos muy serios para las teorías de la indicación, que parecen permitir que una representación (o un estado mental) tenga por contenido prácticamente cualquier cosa. En efecto, si *R* responde por igual a la presencia de la información de que σ como a la presencia de la información de que σ*, entonces la ley causal vigente es «σ V σ*→ S'». ¡Pero, entonces, si #PIRÁMIDE# indica tanto la presencia de pirámides como de cubos, no se deja hueco a la representación errónea! El contenido erróneo es imposible, si la correlación de una representación mental con una disyunción es mejor que con uno cualquiera de los contenido disyuntados. Este es el tan citado Problema de la Disyunción (o del Error) (véase Fodor, 1987, cap. 4; 1990, caps. 2, 3 y 4).

(E_5) *La teoría de la dependencia asimétrica*. Una respuesta a este problema distingue entre conceptos disyuntivos (tales como #GANADO# o #ELECTRODOMÉSTICO#) y conceptos que no lo son (como #VACA#). La respuesta apela a la idea de culminación de un proceso de aprendizaje de la siguiente forma: si el aprendizaje del concepto se culmina habiendo establecido la correlación causal «σ → *R*» y, *después*, la presencia de la información de que σ*, distinta de, activa *R*, entonces *R* representa *erróneamente* la presencia de que σ (Dretske, 1981, 195). Mientras que si el proceso de aprendizaje se culmina habiendo establecido la correlación «σ V σ* → *R*» y, *después*, la presencia de la información de que σ* activa *R*, entonces *R* representa *correctamente* la presencia de que σ*, y no hay error de por medio. Se le ha reprochado a esta maniobra (en Fodor, 1987, cap. 4; 1990, cap. 3) que fija el contenido de una representación *R* por relación a las creencias de alguien más, a saber: a las creencias de quienes controlan que el sujeto *A* culmine correctamente su adquisición

de la representación *R*; y que eso supone remitir la intencionalidad del agente *A* a la de un sujeto diferente. La deseada naturalización del contenido se ve así seriamente obstaculizada.

Se ha citado la solución de Dretske al problema de la disyunción y la crítica consiguiente de Fodor, porque en ese intercambio se aprecian una vez más las tensiones a que da lugar el intento de satisfacer al mismo tiempo la exigencia de reconocer la normatividad del contenido y la de «naturalizarlo». Pero la cosa no acaba aquí. De hecho, el problema de la disyunción tiene raíces más profundas, raíces que penetran por igual en los terrenos de las teorías teleológicas y de las teoría informacionales. El problema real estriba en la robustez del significado (o del contenido), en el hecho de que cualquier representación mental se activa por múltiples causas. Las pirámides activan #PIRÁMIDE#, pero muchas otras cosas que no son pirámides hacen lo mismo. El Louvre activa en usted esa representación sin que #EL LOUVRE# sea sinónima de #PIRÁMIDE#. La información corre por conductos causales abiertos tanto por la selección natural como por el aprendizaje, a pesar de que información no sea lo mismo que contenido (o significado). No todo canal informativo es, así pues, un canal semántico. «Dicho escuetamente, hay a nuestro alrededor mucho menos significado que información» (Fodor, 1990, 93).

La propuesta de Fodor para separar el significado de la información es la siguiente: si bien muchas cosas que no son pirámides (es decir, que carecen de la propiedad de ser pirámides) producen ejemplares particulares de #PIRÁMIDE#, los vínculos causales «no-pirámides → #PIRÁMIDE#» son asimétricamente dependientes del vínculo causal «pirámide → #PIRÁMIDE#». Esto significa que ejemplares de representaciones como #AMENOFIS IV# o como #EL LOUVRE# pueden llevar información sobre pirámides, porque ejemplares de #PIRÁMIDE# llevan información sobre pirámides; pero no a la inversa. No puede haber ejemplares de #PIRÁMIDE# que no estén causados por pirámides sin que haya ejemplares de #PIRÁMIDE# que sí lo estén. No todas las leyes mundo-cerebro (es decir, las relaciones nomológicas entre propiedades del mundo externo y propiedades de eventos en el cerebro) están a la par. Un vínculo nomológico *depende asimétricamente* de otro, si el primero presupone el segundo; es decir, si aquel ejerce su función sólo porque éste se hallaba ya ahí, antes, ejerciendo la suya (Fodor (1990, 96)). Es por ello que *R* significa σ, en lugar de σ^*. Por ello, contenido = información + dependencia asimétrica (Fodor, 1990, 129).

En teorías como las de Dretske y Fodor, la representación se hace posible por la existencia de relaciones causales. Estas relaciones no son, en sí mismas, ni apropiadas ni inapropiadas; no se ajustan a, ni se alejan de, norma algina. Ello no es óbice para que los estados con los contenidos oportunos puedan ejercer funciones en la vida del organismo. Así, para Dretske, las creencias son mapas que nos orientan; si son verdaderas, su

función orientadora se ejerce bien; y si son falsas, mal. Por su parte, Fodor ve los organismos como máquinas de toma de decisiones que actúan tratando de maximizar sus utilidades, lo cual alcanza a suceder cuando sus deseos pueden satisfacerse y cuando sus creencias resultan ser verdaderas. Una vez más, el correcto cometido de una función es el caldo de cultivo de los elementos normativos:

> La evaluación comienza a girar con el resto del engranaje cuando los estados representacionales tienen funciones que se definen por referencia a sus contenidos (cuando un estado que representa el mundo como siendo tal y cual tiene la función de representar el mundo como siendo tal y cual). En estos casos, las representaciones erróneas son fallos de función y, como tales, han de deplorarse" (Fodor, 1990, 129).

Así pues, la función de un estado representacional no determina cuál sea su contenido, pero sí contribuye decisivamente a fijar su valor normativo. Es por esta vía que aproximaciones externalistas como las ilustradas tratan de resolver el problema de la normatividad de lo mental.

VII. CONCLUSION: ¿TIENE DOBLE ASPECTO EL CONTENIDO?

Internalismo y externalismo responden a intuiciones que no es fácil compatibilizar. Para los internalistas, el contenido no depende de nada externo; para los externalistas, sí. Según los primeros, el contenido está constituido por propiedades intrínsecas de la mente. Es por ello que mayoritariamente piensan que sobreviene a las propidades físicas no relacionales del cerebro. Los segundos entienden, por su parte, que el entorno es parte constitutiva del contenido mental y que son las funciones de los estados mentales en la explotación de su medio externo o las conexiones causales que mantengan con éste los factores que fijan ese contenido. Para la gran mayoría de filósofos del momento actual, el internalismo y el externalismo subrayan otros tantos aspectos fundamentales del contenido: el aspecto *interno*, cifrado en el rol causal de la representación o del estado mental, que captura el modo en el que el agente o el organismo ve el mundo y que controla la conducta del primero en el segundo; y el aspecto *externo*, que se identifica con su referente o con sus condiciones de verdad, responsable de las propiedades normativas del estado (o la representación). A cada uno le compete un cometido. El aspecto interno de una representación sería el elemento responsable de su conducta; el externo, el responsable de que esté con el mundo externo en las relaciones que de hecho guarda con él. (Loar ha hablado a este respecto de las relaciones *horizontales* entre representaciaones y de relaciones *verticales* entre representaciones y referentes y/o condiciones de ver-

dad. (Véase Loar, 1981, cap. 2). Ambos ingredientes parecen ser necesarios. El interno da cuenta de por qué el agente actúa como lo hace; el externo de por qué sus acciones son eficaces o cuando menos de por qué sus creencias y deseos están vertidos hacia el mundo [18]. El contenido mental es, entonces, un vector formado por la parte restringida y por la parte amplia (como en la teoría del significado expuesta en Putnam, 1975).

Una pregunta nada trivial —en la cual se cifra el último esfuerzo sistematizador del presente resumen— es la de si esos dos factores o aspectos son independientes; si reflejan preocupaciones distintas y no conectadas entre sí. Una respuesta afirmativa es lo característico de las llamadas teorías del *doble aspecto* (explícitamente asumidas en Field, 1977; Loar, 1981; McGinn, 1982; Lycan, 1985; criticadas en Loewer y LePore, 1987). Una respuesta negativa es lo característico de las opciones descritas hasta el momento. La sugerencia de que ambos aspectos son ineludibles captura una idea (casi) universalmente aceptada. La de que son aspectos independientes goza de un respaldo mucho más limitado. Sin duda, la dificultad de mostrar detalladamente cómo ambos se articulan en una teoría del contenido es real. Quizás se deba a ello el interés con que se han seguido en los últimos tiempos aquellas propuestas (como la de Fodor) que, por considerar el contenido restringido una función de contextos a contenidos amplios, aúnan ambos aspectos de una forma sistemática.

18. Incluso propuestas típicamente externalistas como las de McGinn (1989) y Dretske (1988) parecen limitarse a estados representacionales muy básicos. Ambas obras, en lo que parece ser una maniobra típica, acaban incorporando teorías del rol conceptual para estados mentales con contenidos altamente derivados. La inevitabilidad de aceptar algún contenido restringido es la conclusión de otros autores cuya actitud inicial hacia el externalismo es de simpatía. (Véase Field 1990; Jackson y Pettit, 1993; Sosa, 1993). Una síntesis diferente de ambos factores es la constituida por la semántica del rol conceptual, propugnada por Gilbert Harman. En sus últimas presentaciones, el rol conceptual no sólo incluye las relaciones entre representaciones, percepciones y acciones, sino también los vínculos con objetos y situaciones del mundo externo. (Véase Harman, 1987; 1988).

RESUMEN: CLASIFICACIÓN DE LAS TEORÍAS DEL CONTENIDO MENCIONADAS EN EL ENSAYO

I. Doble aspecto:
   Field, Loar, McGinn, Lycan

II. Único aspecto:
   A. Internalistas
      1. *Estados Globales:*
         (I_1) Loar
      2. *Lenguaje del Pensamiento:*
         a. Holistas: (I_2) Block, Lycan
         b. Atomistas: (I_3) Fodor, LePore
   B. Externalistas
      1. *Anti-individualistas*
         a. (E_1) Kripke, Devitt
         b. (E_2) Burge
      2. *Teleológicas: (E_3) Neander, McGinn, Millikan*
      3. *Causales:*
         a. Estados globales: Stalnaker
         b. Lenguaje del Pensamiento:
            i. (E_4) Dretske
            ii. (E_5) Fodor
      4. *Rol conceptual No-solipsista: Harman*

BIBLIOGRAFÍA

Agar, N. (1993), «What Do Frogs Really Believe?»: *Australasian Journal of Philosophy*, *71*, 1-12.

Almog, J., Perry, J. y Wettstein, H. (comps.) (1989), *Themes from Kaplan*, OUP, Oxford.

Anderson, C. y Owens, J. (comps.) (1990), *Propositional Attitudes*, Chicago University Press, CSLI, Stanford, CA.

Block, N. (1986), «Advertisement for a Semantics for Psychology», en P. French, Th. Uehling, y H. K. Wettstein (comps.), 1986.

Block, N. (1991), «What Narrow Content Is Not», en B. Loewer y G. Rey (comps.), 1991.

Boghossian, P. (1992), «Does an Inferential Role Semantics Rest Upon a Mistake?»: *Mind and Language*, *8*, 27-40.

Boghossian, P. (en prensa), «Analyticity», en C. Wright (comp.) (en prensa).

Brand, M. y Harnish, R. (comps.) (1986), *The Representation of Knowledge and Belief*, The University of Arizona Press, Tucson, Arizona.

Braun, D. (1991), «Content, Causation, and Cognitive Science»: *Australasian Journal of Philosophy*, *69*, 375-389.

Burge, T. (1979), «Individualism and the Mental», en P. French, Th. Uehling y H. Wettstein (comps.), 1979a.
Burge, T. (1981), «Other Bodies», en A. Woodfield (comp.), 1982.
Burge, T. (1986), «Individualism and Psychology»: *Philosophical Review, VC,* 3-46.
Burge, T. (1988), «Cartesian Error and the Objectivity of Perception», en R. H. Grimm y D. D. Merrill (comps.), 1988.
Burge, T. (1989), «Individuation and Causation in Psychology»: *Pacific Philosophical Quarterly, 70*, 303-322.
Brentano, I. (1955/1957), *Psychologie vom empirischen Standpunkte*, Felix Meiner, Hamburg. Traducción inglesa de parte de esta obra en R. M. Chisholm (comp.), 1960.
Chisholm, R. M. (1957), *Perceiving: A Philosophical Study*, Cornell University Press, Ithaca, NY.
Chisholm, R. M. (comp.) (1960), *Realism and the Background of Phenomenology*, Ridgeview, Atascadero, CA.
Churchland, P. M. (1979), *Scientific Realism and The Plasticity of Mind*, Cambridge University Press.
Churchland, P. M. (1981), «Eliminative Materialism and the Propositional Attitudes»: *Journal of Philosophy, 78*, 67-90.
Churchland, P. S. (1980), «Language, Thought, and Information Processing»: *Nous, 14*, 147-170.
Churchland, P. M. y Churchland, P. S. (1983), «Stalking the Wild Epistemic Machine»: *Nous*, 5-18.
Davidson, D. y Harman, G. (comps.), *Words and Objections*, D. Reidel, Dordrecht, Holland.
Dennett, D. (1969), *Content and Consciousness*, Routledge & Kegan Paul, London.
Dennett, D. (1979), *Brainstorms. Philosophical Essays on Mind and Psychology*, The Harvester Press, Hassocks, Sussex.
Dennett, D. (1987), *The Intentional Stance*, The MIT Press, Cambridge, MA.
Devitt, M. (1981), *Designation*, Columbia University Press, New York.
Devitt, M. (1989), «A Narrow Representational Theory of Mind», en S. Silvers (comp.), 1989.
Donnellan, K. (1974), «Speaking of Nothing»: *Philosophical Review, LXXXIII*, 3-31.
Drestke, F. (1981), *Knowledge and the Flow of Information*, Basil Blackwell, Oxford. V. e.: M. Vicedo, M. Guilla y F. Pizarro, *Conocimiento e información*, Salvat, Barcelona, 1987.
Dretske, F. (1988a), *Explaining Behavior*, The MIT Press, Cambridge, MA.
Dretske, F. (1988b), «Reply to Cummins», en R. H. Grimm y D. D. Merrill (comps.), 1988.
Evans, G. (1973), «The Causal Theory of Names»: *Proceedings of the Aristotelian Society*, supl. vol. *XLVII*, 187-208.
Evans, G. (1982), *The Varieties of Reference*, OUP, Oxford.
Field, H. (1977), «Logic, Meaning and Conceptual Role»: *Journal of Philosophy, LXXIV*, 379-408.
Field, H. (1978), «Mental Representation»: *Erkenntnis, 13*, 9-61.

Field, H. (1990), «"Narrow" Aspects of Intentionality and the Information-Theoretic Approach to Content», en E. Villanueva (comp.), 1990.
Fodor, J. (1981), *RePresentations. Philosophical Essays on the Foundations of Cognitive Science*, The Harvester Press, Brighton.
Fodor, J. (1987), *Psychosemantics. The Problem of Meaning in the Philosophy of Mind*, The MIT Press, Cambridge, MA.
Fodor, J. (1990), *A Theory of Content and Other Essays*, The MIT Press, Cambridge, MA.
Fodor, J. (1991a), «A Modal Argument for Narrow Content»: *Journal of Philosophy, LXXXVIII*, 5-26.
Fodor, J. (1991b), «Replies», en B. Loewer y G. Rey (comps.), 1991.
Fodor, J. y LePore, E. (1992), *Holism. A Shopper's Guide*, Blackwell, Oxford.
Frege, G. (1892), «Über Sinn und Bedeutung»: *Zeitschrift für Philosophie und philosophische Kritik, 100*, 25-50. V. e.: U. Moulines, *Estudios sobre semántica*, Ariel, Barcelona, 1971.
French, P, Uehling, Th. y Wettstein, H. K. (comps.) (1979a), *Midwest Studies in Philosophy*, vol. 2: *Contemporary Perspectives in the Philosophy of Language*, University of Minnesota Press, Minneapolis, Minnesota.
French. P, Uehling, Th. y Wettstein, H. K. (comps.) (1979b), *Midwest Studies in Philosphy*, vol. 4: *Studies in Metaphysics*, University of Minnesota Press, Minneapolis, Minnesota.
French. P, Uehling, Th. y Wettstein, H. K. (comps.) (1986), *Midwest Studies in Philosophy*, vol. 10: *Studies in the Philosophy of Mind*, University of Minnesota Press, Minneapolis, Minnesota.
Grimm, R. H. y Merrill, D. D. (comps.) (1988), *Contents of Thought*, University of Arizona Press, Tucson.
Harman, G. (1987), «(Non-Solipsistic) Conceptual Role Semantics», en E. LePore (comp.), 1987.
Harman, G. (1988), «Wide Functionalism», en S. Schiffer y S. Steele (comps.), 1988.
Haugeland, J. (1990), «The Intentionality All Stars», en J. Tomberlin (comp.), 1990.
Jackson, F. y Pettit, P. (1988), «Functionalism and Broad Content»: *Mind, XCVII*, 381-400.
Jackson, F. y Pettit, P. (1993), «Some Content is Narrow», en J. Heil y A. Mele (comps.),1993.
Kaplan, D. (1969), «Quantifying In», en D. Davidson y G. Harman (comps.), 1969.
Kaplan, D. (1989), «Demonstratives: An Essay on the Semantics, Logic, Metaphysics, and Epistemology of Demonstratives and Other Indexicals», en J. Almog, J. Perry y H. Wettstein (comps.), 1989.
Kripke, S. 1980, *Naming and Necessity*, Basil Blackwell, Oxford. V. e.: M. Valdés, *El nombrar y la necesidad*, UNAM, México, 1975.
LePore, E. (comp.) (1987), *New Directions in Semantics*, Academic Press, London.
LePore, E. y Loewer, B. (1987), «Dual Aspect Semantics», en E. LePore (comp.), 1987.
Loar, B. (1981), *Mind and Meaning*, Cambridge University Press.

Loar, B. (1987), «Subjective Intentionality»: *Philosophical Topics, XV*, 89-124.
Loar, B. (1988a), «Social Content and Psychological Content», en R. H. Grimm y D. D. Merrill (comps.), 1988.
Loar, B. (1988b), «Reply: A New Kind of Content», en R. H. Grimm y D. D. Merrill (comps.), 1988.
Loewer, B. y LePore, E. (1987), «Dual Aspect Semantics», en E. LePore (comp.), 1987.
Loewer, B. y Rey, G. (comps.) (1991), *Meaning in Mind: Fodor and His Critics*, Blackwell, Oxford.
Lycan, W. (1985), «The Paradox of Naming», en B. K. Matilal y J. K. Shaw (comps.), 1985.
Lycan, W. (1986), «Thoughts About Things», en M. Brand y R. Harnish (comps.), 1986.
Lycan, W. (1987), *Judgement and Justification*, The MIT Press, Cambridge, MA.
Margalit, A. (comp.) (1979), *Meaning and Use*, Reidel, Dordrecht.
Matilal, B. K. y Shaw, J. L. (comps.) (1985), *Analytical Philosophy in Comparative Perspective*, Reidel, Dordrecht.
McGinn, C. (1982), «The Structure of Content», en A. Woodfield (comp.), 1982.
McGinn, C. (1989), *Mental Content*, Basil Blackwell, Oxford.
Perry, J. (1979), «Frege on Demonstratives»: *Philosophical Review, LXXXVI*, 474-497.
Putnam, H. (1975), *Mind, Language and Reality*, Cambridge University Press. V. e. de J. J. Acero de los ensayos «The meaning of "meaning"» y «¿Es posible la semántica»: *Teorema, XIV/3-4* (1984) 345-405 y *XV/1-2* (1985) 131-145, respectivamente.
Putnam, H. (1978), *Meaning and the Moral Sciences*, Routledge & Kegan Paul, London. V. e.: A. I. Stellino, *El significado y las ciencias morales*, UNAM, México, 1991.
Quine, W. (1951), *From a Logical Point of View*, Harvard University Press, Cambridge, MA. V. e.: J. Sacristán, *Desde un punto de vista lógico*, Ariel, Barcelona, 1962.
Quine, W. (1960), *Word and Object*, The MIT Press, Cambridge, MA. V. e.: J. Sacristán, *Palabra y objeto*, Labor, Barcelona, 1968.
Schiffer, S. (1990), «Fodor's Character», en E. Villanueva (comp.), 1990.
Schiffer, S. y Steele, S. (comps.) (1988), *Cognition and Representation*, Westview, Boulder, Colorado.
Silvers, S. (comp.) (1989), *Rerepresentations: Readings in the Philosophy of Mental Representation*, Dordrecht, Kluwer.
Sosa, E. (1993), «Abilities, Concepts, and Externalism», en J. Heil y A. Mele (comps.),1993.
Stalnaker, R. (1984), *Inquiry*, The MIT Press, Cambridge, MA.
Stalnaker, R. (1989), «On What's in the Head», en J. Tomberlin (comp.), 1989.
Stalnaker, R. (1990), «Narrow Content», en C. Anderson y J. Owens (comp.), 1990.
Stalnaker, R. (1991), «Semantics for the Language of Thought», en B. Loewer y G. Rey (comps.), 1991.

Stampe, D. (1979), «Towards a Causal Theory of Linguistic Representation», en P. French, Th. Uehling y H. Wettstein (comps.), 1979a.
Stich, S. (1983), *From Folk Psychology to Cognitive Science*, The MIT Press, Cambridge, MA.
Tomberlin, J. (1990), *Philosophical Perspectives*, 4: *Action Theory and Philosophy of Mind*, Ridgeview Publishing Company, Atascadero, CA.
Van Gulick, R. (1980), «Functionalism, Information and Content»: *Nature and System, 2*, 139-162.
Van Gulick, R. (1989), «Metaphysical Arguments for Internalism and Why They Don't Work», en S. Silvers (comp.), 1989.
Villanueva, E. (comp.) (1990), *Information, Semantics and Epistemology*, Basil Blackwell, Oxford.
Woodfield, A. (comp.) (1982), *Thought and Object. Essays on Intentionality*, OUP, Oxford.
Wright, C. (comp.) (en prensa), *Blackwell's Companion to the Philosophy of Language*, Basil Blackwell, Oxford.

CAUSALIDAD Y CONTENIDO MENTAL

Manuel Liz

I. INTRODUCCIÓN

Siempre es bueno comenzar con algún tipo de evidencias o, al menos, con algo a partir de lo cual podamos plantear con claridad los problemas. Consideremos los enunciados siguientes:

(1) Tengo un hambre insoportable y no quiero seguir trabajando *porque* no he probado bocado desde el desayuno.
(2) Hemos oído extraños ruidos en la casa *porque* está llena de ratones.
(3) Crees que hay un árbol enfrente *porque* efectivamente hay un árbol enfrente.
(4) Los músculos de mi brazo se contraen *porque* he sentido un fuerte dolor en la mano.
(5) Nos fuimos corriendo a casa; después de pensarlo un rato, llegamos a la conclusión de que habíamos olvidado apagar la plancha y no queríamos provocar un incendio.
(6) Venderá sus acciones *porque* quiere tener dinero en efectivo y cree que vendiendo sus acciones lo conseguirá.
(7) Lo asesinó a tiros *porque* quería matarle y creía que disparándole conseguiría su objetivo.
(8) He sentido una gran alegría *al* volver a ver esas viejas fotos y recordar aquella excursión de pesca.
(9) Decidimos ir a vivir al campo *por la sencilla razón* de que no nos gustaba la ciudad.
(10) Imaginamos que llegarán tarde a la cita *debido a que* suponemos que ayer se acostaron muy tarde y porque de todas formas sabemos que siempre son impuntuales.

(11) Me he asustado y caído de la silla *porque*, mientras estaba pensando en la película de miedo que vi anoche, entraste de repente en la habitación.

A menudo realizamos *explicaciones* como éstas. Son parte de nuestra *psicología natural* (*folk psychology*). La manera habitual de entendernos a nosotros mismos y a los demás como sujetos con una vida mental llena de sensaciones, creencias, deseos, recuerdos, ilusiones, etc., con una vida mental que condiciona nuestras acciones y nuestro modo de relacionarnos con el entorno, hace constantemente uso de tales explicaciones. Esto es un hecho innegable. Y, a primera vista, *parecen explicaciones causales*. Estas explicaciones comparten muchos de los rasgos característicos de cualquier explicación causal. Pueden ser realizadas tanto en primera como en tercera persona (compárese, por ejemplo, 1 con 6), aparentemente pueden llegar a ser bastante objetivas y contrastables (sobre todo cuando, como ocurre en 2 o en 5, la explicación menciona fenómenos mentales compartidos por los sujetos que llevan a cabo tal explicación), pueden ser explicaciones de hechos actuales (véanse 1, 3, 4 y 10) o pasados (véanse 2, 5, 7, 8, 9 y 11), permiten avanzar predicciones de hechos futuros (véase 6), las expresiones con las que formulamos dichas explicaciones pueden ser tan variadas como las que sirven para formular el resto de nuestras explicaciones causales (véanse 8, 9 y 10), etc.

Sabemos que no toda explicación es una explicación causal. Por ejemplo, al explicar por qué no son adecuados ciertos movimientos en el juego del ajedrez, no ofrecemos ninguna explicación causal. Sin embargo, nuestros enunciados de más arriba sí parecen fácilmente *transformables*, acaso sin pérdida alguna de significado, en enunciados directamente causales. Por ejemplo, el enunciado 1 sugiere inmediatamente la reformulación siguiente:

(12) La *causa* de que tenga hambre y de que no quiera seguir trabajando (en las circunstancias en las que me encuentro) es que no he probado bocado desde el desayuno.

Algo semejante cabría hacer con todos los demás enunciados. Las distinciones habituales entre *explanans* y *explanandum*, entre aludir a causas totales o sólo a causas parciales, entre seleccionar unas causas relativas a ciertas condiciones y determinar esas mismas condiciones, etc., serían completamente aplicables aquí. Incluso es posible realizar *generalizaciones en forma de leyes* capaces de referirse a grupos, más o menos significativos o amplios, de esas relaciones causales. Así, por ejemplo, el enunciado

(13) Si S desea A y S cree que haciendo B conseguirá A, entonces, S hará B (*caeteris paribus*)

podría ser una generalización efectuada a partir de casos particulares como 6 o 7, generalización cuya cláusula *caeteris paribus* resaltaría explícitamente el carácter *relativo a ciertas condiciones* de las relaciones causales pretendidamente aludidas en tales casos particulares. Sin esa cláusula, 13 expresaría algo demasiado fuerte, algo que muchas veces es falso. Su cláusula *caeteris paribus* nos previene contra todas esas excepciones indicando que, a pesar de ellas, existe un grupo relevante de casos sobre los que es posible y explicativamente provechoso generalizar. El mismo efecto podría haberse conseguido añadiendo condiciones al antecedente o restringiendo el alcance del consecuente. Sin embargo, la introducción de cláusulas *caeteris paribus* resulta mucho más directa. El enunciado 13 sería, pues, un perfecto candidato al estatus de *ley psico-física* capaz de subsumir nómicamente, con la ayuda de una cláusula *caeteris paribus*, casos como los ofrecidos en 6 o 7. Y por razones similares, enunciados como 8, 9 o 10 podrían generalizarse y dar lugar a leyes *caeteris paribus* puramente *psicológicas*.

De cualquier forma, nos encontramos ante una variedad sumamente peculiar de explicaciones causales. En ellas se mencionan siempre ciertos fenómenos mentales:

A) *como efecto de algo no mental* (véase 1, 2 y 3),
B) *como causa de algo no mental* (véase 4, 5, 6 y 7),
C) *o como causa a la vez que como efecto* (véase 8, 9, 10 y 11).

Estaríamos, en consecuencia, ante la posibilidad de que ciertas relaciones causales involucren fenómenos mentales por el lado de los efectos, por el lado de las causas o por ambos lados. Y esto ya no parece tan fácil de entender. ¿Cómo pueden llegar a producirse causalmente cosas del tipo de nuestras sensaciones, deseos o creencias? ¿Cómo es posible que algo mental intervenga causalmente en los procesos desarrollados en el mundo natural? *Una parte* de nuestra autoconcepción como personas se inclina a establecer un radical contraste entre nuestra vida mental, con todos sus productos y manifestaciones, y el mundo natural. El anti-reduccionismo respecto a lo mental es la versión filosófica de esta inclinación. Pero, como estamos señalando, *otra parte* de esa misma autoconcepción siente una no menos fuerte inclinación a vernos causalmente involucrados en el curso natural de las cosas. ¿Cómo entender esta polaridad?

Ninguna de estas preguntas tiene ya una respuesta sencilla o inmediata. Sin embargo, son las preguntas que aquí nos vamos a hacer. Más concretamente, el *objetivo* de este trabajo es explorar, en primer lugar,

hasta qué punto algunas de esas explicaciones en las que se apela a fenómenos mentales realmente pueden ser explicaciones causales y, en segundo lugar, hasta qué punto esas explicaciones causales serían capaces de identificar y describir genuinas relaciones causales en las que intervengan fenómenos mentales.

Pero volvamos a los enunciados de más arriba. Ya hemos dicho que lo mental interviene en estas explicaciones como causa, como efecto o a la vez como causa y efecto. Ciertas *propiedades mentales* permiten identificar esos fenómenos mentales que se mencionan. Los fenómenos mentales pueden ser *estados mentales* (en 1, tener un hambre insoportable y no querer seguir trabajando; en 2, oír ruidos en la casa, etc.) o *procesos mentales* (en 5, llegar a la conclusión, después de pensarlo un rato, de haber olvidado apagar la plancha, etc.). Estos procesos mentales serían transiciones de unos estados mentales a otros. Y los estados mentales pueden ser tanto *intencionales* (en 1, no querer seguir trabajando; en 3, creer que hay un árbol enfrente, etc.), estados con un peculiar *contenido semántico* (contenido que es proposicionalmente expresado en las sentencias subordinadas «seguir trabajando», «hay un árbol enfrente», etc.), como *no intencionales* (en 1, tener un hambre insoportable; en 2, oír extraños ruidos en la casa; en 4, sentir un fuerte dolor en la mano, etc.), con un especial *contenido cualitativo o fenoménico* (contenido tal vez últimamente inefable, pero aludido muchas veces mediante adverbios o adjetivos que, como en nuestros ejemplos, cualifican el tipo de estado de que se trata).

Algunas de las explicaciones que estamos analizando mencionan fenómenos mentales como *causa de acciones* [1] (casos del tipo B). Con ello se supone que, de alguna manera, puede trazarse una distinción entre las cosas que *nos pasan* y las cosas que *hacemos*. Las cosas que hacemos obedecen a *razones*. Las creencias y los deseos serían los dos tipos de razones más comunes de nuestra acción. Pero una cosa son las *razones que se pueden tener para hacer algo* y otra cosa son las *razones por las que efectivamente se hace algo*. Las primeras sólo racionalizan una acción, las segundas además la explican. Y ¿cómo pueden las razones explicar una acción y no sólo racionalizarla? La respuesta que recientemente dio Davidson [2] a esta pregunta ha sido tremendamente influyente. Las explica-

1. Cabe argumentar que toda acción colectiva, por compleja y sofisticada que sea, está basada en acciones individuales. En contextos diferentes, puede verse argumentada esta idea, por ejemplo, en Mosterin (1978), Quintanilla (1989) y Tuomela (1989). Curiosamente, algunos autores sitúan el conocimiento de la propia acción intencional en el núcleo mismo de nuestro concepto de causalidad. ¿Qué concepto de causalidad, se viene a preguntar aquí, podría tener un ser capaz de tener percepciones y pensamientos, pero absolutamente incapaz siquiera de pestañear o de dirigir su vista hacia determinado lugar? No sé muy bien dónde nos llevaría discutir aquí con detalle esta cuestión. La dejo, pues, simplemente indicada. Sobre este tema, véase Alvarez (1990).

2. Davison (1963).

ciones de la acción en base a razones han de ser explicaciones *causales*. Las razones explican las acciones siendo sus causas. De acuerdo a esto, la acción intencional consistiría en un tipo muy especial de relación causal en la cual parte de las condiciones causales productoras del efecto son cosas como creencias y deseos. Pero, al margen de los planteamientos concretos de Davidson, la idea de que las razones son causa de la acción no sólo pertenece a nuestra más profunda *folk psychology*, sino que ha estado siempre presente en la filosofía desde sus mismos comienzos, y podemos así encontrar alusiones explícitas a ella en autores como, por ejemplo, Platón o Aristóteles [3].

Otras explicaciones causales que mencionan fenómenos mentales los hacen intervenir como un *efecto causal*. Lo mental aparece a veces como un producto causal de fenómenos no mentales (casos del tipo A) o como un producto causal de otros fenómenos que también son mentales (casos del tipo C). Nuestras sensaciones y percepciones, los fenómenos estudiados por las teorías de la psicomotricidad, la mayor parte del conocimiento empírico y práctico, la experimentación, etc., serían difícilmente explicables si no admitiéramos que algunos fenómenos mentales pueden ser un efecto causal de fenómenos no mentales. Nuestros razonamientos, lo que solemos llamar «el curso de nuestros pensamientos», también invitan a ser explicados a través de relaciones causales en las que unos fenómenos mentales son producidos por otros fenómenos ahora también mentales.

Estamos considerando explicaciones causales en las que siempre interviene algo mental. Y esto las hace problemáticas. Pero muchas veces también se mencionan en ellas *fenómenos no mentales* como *causa* (los mismos casos del tipo A), como *efecto* (los mismos casos del tipo B), o a la vez como *causa y efecto* (véase, por ejemplo, el caso 11). Son, por lo demás, numerosos los casos en los que se mencionan fenómenos tanto mentales como no mentales *simultáneamente* en la causa y en el efecto (véase asimismo 11). Ciertas *propiedades físicas* nos permiten identificar estos fenómenos no mentales que podrían ser, también aquí, *estados* de cosas no mentales o *procesos* no mentales, siendo estos procesos transiciones de unos estados no mentales a otros.

Los estados o procesos no mentales son, en un sentido muy general, estados o procesos *físicos*. Es un tremendo problema determinar con precisión en qué consisten los contenidos intencionales semánticos y los contenidos cualitativos que caracterizaban a los fenómenos mentales. Pero no lo es menos determinar con precisión hasta dónde puede llegar

3. Véase Platón, *Fedón*, 98c-99a, y Aristóteles, *Ética a Nicómaco*, Libro 6, 1139a. Otra idea, muy relacionada con ésta, que asimismo ha estado presente en la filosofía desde su mismo origen, es la de que algunas propiedades semánticas de esas razones (por ejemplo, la verdad de las creencias) también tienen relevancia explicativa y eficacia causal (por ejemplo, en relación al éxito de una acción). No abordaremos aquí, sin embargo, estas cuestiones. Sobre este tema, véase Broncano (1993).

este sentido *general* de lo físico que nos sirve para caracterizar ahora los fenómenos no mentales. La contracción de los músculos de mi brazo (véase 4) parece ser un proceso constituido únicamente por estados de cosas indudablemente físicas. Tal vez también sean físicas cosas como no haber probado bocado desde el desayuno, que la casa esté llena de ratones, etc. (estados o procesos no mentales como los mencionados, por ejemplo, en 1, 3, 4, 5 y 11). Sin embargo, vender las acciones o asesinar a tiros (véanse 6 y 7) no parecen ya estados que podamos tipificar sin problemas como físicos.

Seguramente sea aquí pertinente cierta distinción entre *conducta* y *acción*, entendiendo por conducta algo mucho más objetivamente caracterizable y cercano a lo físico que la acción. Y acaso fuera también conveniente intentar trazar una distinción entre lo *físico* y otras cosas aún no mentales pero tampoco puramente físicas, cosas que podríamos llamar *materiales* [4]. De todas formas, hasta cierto punto esto sería por ahora una cuestión secundaria. Si se concluyera que estados como vender las acciones o asesinar a tiros son estados en parte *también* mentales, sólo ocurriría que las explicaciones que los mencionan resultarían tener *más* componentes mentales de lo que parecía.

Son muchas las mezclas posibles de todos los elementos que acabamos de indicar. Ciertamente, las explicaciones que mencionan fenómenos mentales forman una clase bastante *heterogénea*. Y el caso es que tal vez las diferencias sean mucho mayores que las similaridades. Tal vez no sea demasiado provechoso, después de todo, pensar que nos encontramos con un *único* tipo de explicaciones causales y abordar los problemas desde este punto de vista. Pero no nos volvamos pesimistas.

II. RELEVANCIA EXPLICATIVA Y EFICACIA CAUSAL DE LO MENTAL

En toda explicación causal se intenta siempre, de una u otra forma, seguir el rastro de posibles relaciones causales que tengan lugar en la realidad. La apelación a fenómenos mentales sería *explicativamente relevante*, en este sentido, si nos pone en la pista de alguna relación causal interesante. Y los propios fenómenos mentales que son mencionados en tales explicaciones serían *causalmente eficaces* si consiguen encadenarse apropiadamente en esas relaciones causales. Pero la relevancia explicativa es una cuestión epistémica, y la eficacia causal es una cuestión ontológica.

Resulta difícil imaginar cómo podría llegar a ser razonable admitir la eficacia causal de lo mental si la apelación a esos fenómenos mentales no

4. Acerca de los muchos problemas que surgen en cuanto queremos definir con mayor precisión los propios conceptos de lo «físico», lo «material» o lo «natural», véanse Esquivel (1982), Liz (1993) y Vázquez y Liz (1993).

pudiera ser *nunca* explicativamente relevante. Pero no lo es tanto imaginar una relevancia explicativa de lo mental *sin* que lo mental pudiera llegar a ser causalmente eficaz. La apelación a fenómenos mentales puede ponernos sobre la pista de *ciertas* relaciones causales en las que, sin embargo, esos fenómenos mentales no intervengan. Este problema general se agrava cuando se conecta con el alcance sumamente restringido a ciertas condiciones que, como hemos visto, tienen las explicaciones y leyes causales que mencionan fenómenos mentales. Su relevancia explicativa podría, entonces, depender justamente de esas *circunstancias particulares* que se suponen fijas sin que tuviera que haber realmente *ningún* fenómeno mental causalmente eficaz. Esta situación no es nada extraña. Pensemos, por ejemplo, en la siguiente explicación:

(14) Este caballo será muy veloz porque sus padres fueron grandes campeones.

Una generalización sobre casos parecidos daría lugar a cierta ley *caeteris paribus* como ésta:

(15) Si el fenotipo de cierta pareja de individuos de una determinada especie es X, entonces el fenotipo de su inmediata descendencia será también X (*caeteris paribus*).

Ciertamente, se trata de una ley aplicable a numerosos casos. Su relevancia explicativa en muchos contextos es indudable. Y, sin embargo, sabemos que los fenotipos de los padres *carecen realmente* de eficacia causal en relación a los fenotipos de los hijos. Tenemos aquí, pues, una relevancia explicativa *sin* ninguna eficacia causal asociada. Nuestra ley recoge cierta regularidad sin contrapartida directa con ninguna relación causal que responda a los términos de esa regularidad. Ciertamente, existe una serie de complejas relaciones causales subyacentes que explican esa regularidad. Y mientras se mantengan fijas esas relaciones causales, se mantendrá la regularidad. En nuestro ejemplo, la cláusula *caeteris paribus* asegura que se mantienen fijas esas relaciones. Las cláusulas *caeteris paribus a veces* funcionan así. Pero no siempre. *Otras veces* consiguen realmente delimitar variedades un tanto especiales de genuinas relaciones causales que responden a los términos en los que se plantea la regularidad, *obteniendose* en tal caso *tanto* una relevancia explicativa *como* una eficacia causal. Esto es lo que ocurre cuando, por ejemplo, decimos

(16) Si aumentamos la temperatura de cierta sustancia más allá de cierto punto crítico, entonces se producirá una explosión (*caeteris paribus*).

Aquí sí que parece más claro que existirían genuinas relaciones causales que, bajo las restricciones impuestas por la cláusula *caeteris paribus*, son *detectadas* por los términos en los que expresamos nuestra ley. En el contexto de las explicaciones causales que mencionan fenómenos mentales, el problema consiste en *determinar cuál* de estos papeles desempeñan las cláusulas *caeteris paribus* que aparecen en ellas, y en las leyes que las subsumen. En el segundo caso, tendremos eficacia causal además de relevancia explicativa; en el primero, no.

Otro importante problema, ahora tanto para la eficacia causal de lo mental como para su relevancia explicativa, tiene que ver con las relaciones *conceptuales* que en ocasiones encontramos en las explicaciones que mencionan fenómenos mentales. Para que una explicación consiga captar relaciones causales genuinas, una condición mínima exigible es que entre el *explanans* y el *explanandum no* exista una ligazón tan estrecha como para que el segundo implique conceptualmente al primero. Pero esto es justamente lo que pasaría si fuera una *verdad conceptual* que, por ejemplo, asesinar a tiros necesariamente implicara querer matar a alguien y creer que disparándole se conseguirá ese objetivo. La explicación 7 no podría ser, entonces, considerada una auténtica explicación causal. Algo fallaría. Estas explicaciones se parecen más a las explicaciones que hacemos de la adecuación o no de ciertos movimientos en el juego del ajedrez que a explicaciones que intenten detectar genuinas relaciones causales. Y una duda que repetidamente surge en este ámbito es la de si no pasará esto mismo con *gran parte* de las explicaciones pretendidamente causales que mencionan fenómenos mentales. Especialmente, con todas aquellas en las que se explica la producción de una acción en base a sus razones y con todas aquellas en las que se explican unos fenómenos mentales en base a otros fenómenos mentales que los justifican y hacen razonables. Ahora bien, ¿cómo *saber* si realmente existe o no tal nexo conceptual? La respuesta de Davidson[5] vuelve a ser aquí pertinente, casi inevitable. Ese nexo conceptual no existirá si los fenómenos, mentales o no mentales, mencionados en nuestros *explanans realmente causan* los fenómenos, mentales o no mentales, mencionados en nuestros *explananda*. Nuevamente, una relevancia explicativa sin algún tipo de eficacia causal parece algo lleno de problemas

III. LA EFICACIA CAUSAL DE LO MENTAL FRENTE A LA CLAUSURA CAUSAL DE LO FÍSICO

Pero, ¿cómo puede ser causalmente eficaz algo mental? ¿Cómo puede algo mental intervenir en una cadena causal? En la ciencia se acepta

5. Nos referimos, sobre todo, a Davison (1963).

cierto principio de *clausura causal del mundo físico* que impide la aceptación directa de la eficacia causal de lo mental. Según este principio, lo físico sólo podría tener causas físicas. Si algo físico tiene una causa, esa causa ha de ser también física. Es imprescindible *reorganizar* de alguna manera nuestros conceptos para conseguir conciliar el anterior principio con la idea de que algunas de las explicaciones y leyes que recogíamos antes realmente consiguen detectar o describir genuinas relaciones causales. Veamos a continuación algunas de estas posibles reorganizaciones.

1. *Epifenomenalismo*

La clausura causal de lo físico, tal como la hemos presentado, sería compatible con el carácter epifenoménico de lo mental. Según el epifenomenalismo, los fenómenos mentales son causados por fenómenos físicos (por fenómenos neurales, en nuestro caso biológico), pero carecen por sí mismos de cualquier eficacia causal. Son causados, pero no causan nada. No lo hacen, al menos, en cuanto tales fenómenos mentales. Los fenómenos mentales son tan inertes causalmente como las imágenes de un espejo que reflejen cierto proceso causal, como los síntomas de una enfermedad en relación a su evolución posterior o como los fotogramas de una película en relación al desarrollo de la misma. Para el epifenomenalismo, no todas las aparentes explicaciones y leyes que mencionan fenómenos mentales denotan auténticas conexiones causales. Sólo algunas del *tipo A* conseguirían hacerlo.

Si hubiéramos formulado el principio de clausura causal de lo físico diciendo que lo físico sólo puede tener causas físicas y efectos también físicos, el epifenomenalismo no tendría cabida. Pero mientras que nuestro principio sí parece jugar un papel importante en la ciencia, al menos un papel metodológico y regulativo, este nuevo principio introduciría un elemento irrelevante. ¿Qué importancia tendría que lo físico pudiera llegar a producir algo no físico si nada que no sea físico puede ser causalmente eficaz? El epifenomenalismo se hace hueco a través de esta irrelevancia. Sin embargo, por esto mismo, el epifenomenalismo nos priva de gran parte de lo que se intentaba asegurar. Nos priva de la eficacia causal de lo mental cuando nuestra acción obedece a nuestras creencias y deseos (casos del tipo B), y de la eficacia causal de lo mental a través del curso de nuestros pensamientos (casos del tipo C). Para algunos, esto supondría la reducción al absurdo del epifenomenalismo. Pero, aunque no fuera así, es indudable que una realidad sólo epifenomenalista para lo mental es una realidad bastante brumosa. Lo mental sería sólo un subproducto, algo así como el sonido de la sirena de vapor que acompaña al funcionamiento de una locomotora sin infuir para nada en su maquinaria [6].

6. El símil es de Thomas Huxley, uno de los máximos exponentes del epifenomenalismo moderno. Véase Huley (1898). Podemos encontrar una apasionada y muy interesante crítica al epifeno-

2. *Interaccionismos no sustancialistas: el emergentismo de propiedades y el monismo anómalo de Donald Davidson*

El interaccionismo es una de las posiciones más clásicas en este terreno. Es, además, la posición ordinariamente más frecuente. Descartes fue un interaccionista. Pero su interaccionismo [7] era *sustancialista*, se desarrollaba en el marco de un dualismo de sustancias. La mente y la materia eran dos sustancias completamente diferentes. Pese a ello, según Descartes, podían interactuar a través de nuestros cerebros, concretamente a través de la glándula pineal. El problema, muchas veces señalado, consistía en que algo inextenso y sin partes, como era la mente, difícilmente podía interactuar con algo material que es extenso y tiene partes. La interacción sustancialista cartesiana era, en el mejor de los casos, un misterio. Un misterio que, además, hacía peligrar el anterior principio de clausura causal de lo físico. Todos los interaccionismos sustancialistas se encuentran con este problema. Así, el interaccionismo sustancialista recientemente propuesto por Karl Popper y John Eccles [8], por ejemplo, sigue intentando localizar «en algún lugar del cerebro» el foco de tal interacción imposible. Pero hay otras *formas no sustancialistas de interaccionismo* que sí se esfuerzan en respetar la clausura causal de lo físico. Examinaremos dos: cierto emergentismo no sustancialista y el monismo anómalo de Davidson. En ellas, la interacción de lo mental con lo físico se produce o bien a través de ciertas propiedades muy especiales, en el primer caso, o bien entre eventos, en el segundo caso, intentando siempre no salir fuera del marco de relaciones causales impuesto por lo físico.

El *emergentismo* es una posición habitualmente ligada al vitalismo y al problema del surgimiento de la vida, de las peculiares propiedades biológicas, tal como fue tematizado a principios de siglo. Las tesis emergentistas, sin embargo, resultan muy fácilmente exportables también al terreno de la vida mental. De hecho, la mayoría de los autores emergentistas clásicos así lo hicieron [9]. El emergentismo no admite otro tipo de *entidades* aparte de las entidades físicas. Pero afirma que ciertos niveles de complejidad física originan la existencia de *propiedades emergentes*, propiedades no reducibles a las propiedades físicas pero que imprimen una peculiar eficacia causal a aquellos sistemas que las posean.

En cierto sentido, el emergentismo sería compatible con el principio de clausura causal de lo físico. Todo lo que puede ser una causa sigue

menalismo en Burge (1993). Para este autor, el papel del epifenomenalismo en filosofía de la mente sería análogo al papel jugado por el esceptitismo en epistemología.

7. Véanse, en particular, sus obras *Meditaciones metafísicas* y *Las pasiones del alma*.
8. Popper y Eccles (1977).
9. Véase, por ejemplo, Broad (1925) y Morgan (1923).

siendo una entidad física. Sin embargo, hay dos cosas que nunca han estado suficientemente claras en el emergentismo: 1) ¿Cómo entender el surgimiento de propiedades emergentes? ¿Es ese mismo surgimiento una relación causal? ¿qué otro tipo de relación podría ser? y 2) ¿Cómo entender la eficacia causal de las propiedades emergentes? ¿Depende de ciertas propiedades físicas subyacentes? ¿De qué tipo de dependencia se trataría? Estas *ambigüedades* son graves. Si entendemos, por ejemplo, el surgimiento de las propiedades emergentes como algo causal y si hacemos depender muy directamente su eficacia causal de ciertas propiedades físicas, lo que obtenemos simplemente es un epifenomenalismo. Pero si, por otra parte, intentamos escapar del epifenomenalismo, concediendo mayor autonomía causal a las propiedades emergentes, resulta difícil ver cómo se conseguiría ahora algún grado de dependencia apropiado respecto de sus sustratos físicos capaz de evitar un interaccionismo sustancialista. El emergentismo es, realmente, una posición poco elaborada en sus detalles.

Uno de los autores actualmente más cercanos al emergentismo es John Searle [10]. La conciencia, según él, es una propiedad causalmente emergente de ciertos sistemas complejos. No puede ser explicada sólo a partir de la composición física de tales sistemas, pero sí puede ser explicada teniendo en cuenta las interacciones causales entre sus componentes físicos. La conciencia es una propiedad emergente de ciertos sistemas de neuronas, de la misma forma que estar en estado sólido o líquido son propiedades emergentes de ciertos sistemas de moléculas [11]. Mientras no se hagan demasiadas preguntas, todo parece ir bien. Sin embargo, los dos problemas que hemos señalado más arriba siguen siendo, también para Searle, obstáculos insalvables.

Vayamos al *monismo anómalo* de Davidson. Su influencia ha sido ciertamente enorme en la filosofía de la mente de las últimas dos décadas. Y por ello, le dedicaremos una especial atención. Comencemos introduciendo su noción de *evento*. Hasta ahora, hemos estado hablando de fenómenos mentales y no mentales que podían ser tanto estados como procesos. Pues bien, un evento sería un peculiar estado o proceso que tiene lugar, un acontecimiento concreto, algo particular que ocurre. Los eventos son de un tipo u otro. Podemos *clasificar* y *describir* los eventos, incluso el mismo evento, de muchas formas [12]. Según Davidson [13], los eventos serían mentales o físicos según puedan ser descritos como perte-

10. Véase Searle (1980, 1984 y 1992).

11. Véase especialmente Searle (1992; cap.5).

12. De la extensa literatura acerca del tema de los eventos, quisiera resaltar los recientes trabajos de Bennett (1988) y Lombard (1986).

13. Véase especialmente Davison (1970, 1973, 1974 y 1993). Un interesante análisis crítico del monismo anómalo se encuentra en Quesada (1984).

necientes, respectivamente, a *tipos* caracterizables mediante propiedades mentales o físicas.

El monismo anómalo de Davidson admite *interacciones causales* en las que intervengan eventos mentales. Los eventos mentales pueden intervenir en los procesos causales al ser *idénticos* a ciertos eventos físicos. Pero rechaza, sin paliativos, que se puedan *reducir* las propiedades mentales a propiedades físicas a través de definiciones o de leyes estrictas. Lo primero hace del monismo anómalo justamente una variedad de *monismo fisicalista* que admite la intervención causal de los eventos mentales. Lo segundo lo convierte en *anómalo* respecto a la manera como puede realizarse esa intervención. Lo mental acaso pueda conectarse con lo físico de otra forma, pero no a través de definiciones ni de leyes estrictas.

Necesitamos explicar todo esto un poco más. Davidson mantiene un enfoque humeano de la causalidad como *regularidad nómica*: si un evento particular causa otro evento particular, entonces esos eventos deben ser de unos tipos tales que *exista* una *ley adecuada* capaz de conectar causalmente los eventos del primer tipo con los del segundo. Un caso sumamente importante de ley adecuada sería el de una generalización verdadera, tan determinista como pueda serlo la naturaleza, sin excepciones ni cláusulas *caeteris paribus* y que trate al universo como un sistema cerrado. Una ley con estas características sería lo que Davidson llama una *ley estricta* [14]. Y únicamente en la física sería posible encontrar leyes de este tipo.

Davidson sostiene que no es posible reducir definicionalmente las propiedades mentales a propiedades físicas y que las leyes psico-físicas, ya sean reductivas o no, o las leyes puramente psicológicas *no* pueden ser nunca leyes estrictas [15]. La adscripción de fenómenos mentales, especialmente de estados mentales intencionales, obedece *constitutivamente* a ciertas constricciones de *racionalidad*. Intentamos siempre salvar la consistencia y la completud de la vida mental de los sujetos a los que adscribimos fenómenos mentales, así como la armonía con sus historias pasadas y sus entornos. Y tales constricciones de racionalidad llenan de excepciones esas adscripciones y las hacen depender siempre de problemáticas cláusulas *caeteris paribus*. El único camino para independizarnos realmente de ellas pasaría por una completa reducción fisicalista de la misma racionalidad. Lo cual es ya mucho pedir [16].

De todas formas, insiste Davidson, en este tema de la causalidad habría que distinguir muy bien entre las relaciones causales mismas y las explicaciones y leyes causales. La concepción humeana de la causalidad

14. Véase Davison (1970 y 1993).

15. Davison (1970) discute el caso particular de las leyes puramente psicológicas.

16. Este proyecto de naturalizar la misma razón ha resultado ser, no obstante, uno de los motores más potentes de la filosofía de nuestro siglo. Véase Ezquerro (1984).

como regularidad nómica pone en conexión ambas cosas, afirmando que siempre que se dé una relación causal ha de existir una ley causal adecuada. En otras palabras, que no hay relaciones causales sin leyes. Pero las relaciones causales son relaciones *diádicas* puramente *extensionales*. Lo que se relaciona son simplemente pares de eventos. Únicamente las explicaciones y leyes causales hacen intervenir propiedades y tipificaciones de eventos. Por ello, Davidson no vería ningún problema en el *principio de clausura causal del mundo físico*. En el contexto de una concepción humeana de la causalidad, tal principio implica la existencia de leyes físicas adecuadas. Y ya hemos visto que, para Davidson, las estrictas leyes de la física serían un caso claro de leyes adecuadas. Pero nada de esto tiene consecuencias excluyentes respecto a la posible existencia de otras *explicaciones y leyes causales también adecuadas*. La clausura causal del mundo físico es *compatible* con la existencia de adecuadas leyes causales no estrictas ni físicas que mencionen fenómenos mentales tipificando de otra forma los mismos eventos. Al fin y al cabo, las leyes y las explicaciones son también *relativas* a unos intereses. Y hay muchos intereses no fisicalistas [17].

Han sido abundantes las *críticas* a la posición de Davidson. Y también han sido abundantes las respuestas y matizaciones. Examinaremos sólo algunas. Jaegwon Kim ha señalado, por ejemplo, que si entendemos un evento como la ejemplificación por un objeto de una propiedad en un tiempo determinado, y si se afirma que todo evento mental es también un evento físico, entonces esta identidad de eventos debería implicar alguna identidad entre las propiedades mentales que los tipifican y ciertas propiedades físicas [18]. Algún tipo de *definición o de ley psico-física estricta* debería existir. Lo mental no sería ya tan anómalo. Davidson escapa de este problema recordando que no deben confundirse los *eventos concretos*, entidades particulares, con los *tipos de eventos*. Un mismo evento concreto puede ejemplificar propiedades muy distintas, puede ser tipificado de muchas formas. Y la identidad entre eventos concretos sólo implicaría, en cualquier caso, que esos eventos ejemplifican exactamente las mismas propiedades, todas ellas, no que tengan que ser idénticas entre sí *algunas* de esas propiedades.

Pero los eventos concretos de Davidson son muy curiosos. Deben ser considerados entidades tales que han de tener todas sus propiedades *esencialmente*. Esto resulta especialmente claro si examinamos la respuesta de Davidson al siguiente problema planteado por Ernesto Sosa [19].

17. Véase, en particular, Davison (1993: 15-6).
18. Véase, por ejemplo, Kim (1978, 1979 y 1989a).
19. Reconstruimos aquí el caso presentado por Sosa (1984: 278). Pueden encontrarse ejemplos similares en Dretske (1989), Honderich (1982), Achestein (1979), Stouland (1976) y Johnston (1985); y una discusión de los mismos en LePore y Loewer (1987).

Alguien es asesinado mediante un disparo ruidoso. El carácter ruidoso del disparo es irrelevante respecto a la producción causal de la muerte por el disparo. *Si la pistola hubiera tenido un silenciador, el disparo habría matado igualmente a la víctima.* De la misma forma, en el monismo anómalo, las propiedades mentales que tienen los eventos mentales son causalmente irrelevantes respecto a los efectos que esos eventos puedan causar. Aunque los eventos mentales no tuvieran las propiedades mentales que tienen, o aunque tuvieran otras, seguirían teniendo los mismos efectos.

Según Davidson[20], el problema radica en que el contrafáctico que hemos enfatizado en el anterior párrafo es tremendamente *ambiguo*. Parece como si, en él, la descripción «el disparo» se refiriera al primer disparo. Sin embargo, esto sería contradictorio, pues un mismo disparo no puede ser a la vez ruidoso y silencioso. Si, por el contrario, se refiere al disparo silencioso, seguramente sea cierto que mataría igualmente a la víctima. Pero, entonces, afirma Davidson, ya no sería la *misma* muerte. Según Davidson, los efectos causales de los eventos mentales no podrían ser los mismos si esos eventos tuvieran propiedades distintas a las que tienen. Y esto le compromete con un tremendo *esencialismo*. Cada evento concreto ha de tener todas sus propiedades esencialmente. Una muerte provocada por un disparo ruidoso, por ejemplo, nunca podría ser la *misma* muerte que una muerte producida por un disparo silencioso.

En cualquier caso, y manteniendo el anomalismo, los eventos mentales resultan, en sentido estricto, causalmente inertes justamente en cuanto mentales. La *estricta* eficacia causal de lo *mental*, en el monismo anómalo de Davidson, consiste únicamente en la *estricta* eficacia causal de lo *físico*. Las propiedades mentales no suponen *estrictamente* ninguna diferencia causal. Es cierto, como acabamos de ver, que puede encontrarse un sentido en el que si cambian las propiedades mentales, pueden cambiar los efectos, pero sólo porque, al tener todas sus propiedades esencialmente, siempre cambiarían también los eventos concretos que actúan como causas. Sin embargo, como acertadamente señala E. Sosa[21], lo mental en este sentido tendría *tanta* eficacia causal como la puede tener el hecho de que en la pistola de más arriba esté o no esté cualquier simple mota de polvo. Y esto, ciertamente, *no es mucha* eficacia causal. No es, al menos, el tipo de eficacia causal que se buscaba. Estaríamos, pues, ante algo así como un *epifenomenalismo* respecto a las propiedades mentales[22].

¿Cómo escapar de tal epifenomenalismo? ¿Cómo hacer que las propiedades mentales sean, después de todo, causalmente eficaces? Se nos

20. Véase Davison (1993).

21. Véase Sosa (1993).

22. Han acusado de epifenomenalismo al monismo anómalo, entre otros, Dretske (1989), Fodor (1989), Honderich (1982), Johnston (1985), Kim (1984c, 1989a, 1993a y 1993b), Sosa (1984 y 1993) y Stoutland (1985). Aparte de Davison (1993), una sofisticada defensa frente a tal acusación puede encontrarse en McLaughlin (1983).

presentan *tres* opciones básicas: 1) consentir que leyes no estrictas, leyes *caeteris paribus* por ejemplo, puedan detectar y subsumir adecuadamente genuinas relaciones causales, 2) admitir algún tipo de concepción no humeana de la causalidad y 3) la propia alternativa preferida por Davidson consistente en recurrir a cierta versión débil de la noción de sobreveniencia.

Ya hemos indicado que el mismo Davidson admite la posibilidad de considerar como *leyes adecuadas*, capaces de subsumir nómicamente relaciones causales genuinas, ciertas *leyes no estrictas* y que, en cualquier caso, esto sería *compatible* con su monismo anómalo. Davidson nos recuerda, entre otros, los argumentos de Fodor a favor de la eficacia causal de lo mental en base a la existencia de obvias regularidades psicológicas y psico-físicas[23]. Todas las ciencias especiales estarían llenas de tales leyes no estrictas; concretamente, de leyes *caeteris paribus*. Nuestro anterior enunciado 13 sería un ejemplo de este tipo de leyes. Dedicaremos, más adelante, un apartado especial a analizar con detalle esta perspectiva.

La segunda opción ha sido explotada por bastantes autores, llevándose a cabo generalmente en base a análisis *contrafácticos* de la causalidad[24]. Según estos enfoques, un evento c causaría otro evento e en el caso de que sea verdad que si c no hubiera ocurrido, e tampoco habría ocurrido. Las causas serían condiciones necesarias para los efectos. Podría darse así, se piensa, una relación causal *sin* que existieran leyes que conectasen los tipos de eventos sobre los que se establece tal relación causal. El principal problema con esta alternativa es que resulta muy difícil analizar la verdad de sus enunciados contrafácticos, acudiendo por ejemplo a una semántica de mundos posibles, sin depender aquí de algún tipo de *leyes*.

La tercera opción tampoco está exenta de problemas. Davidson sostiene que el monismo anómalo no implica ningún epifenomenalismo. Pero reconoce que es *compatible* con él. A fin de garantizar que las propiedades mentales entrañen algún tipo de eficacia causal, suplementa su monismo anómalo con cierta tesis de la *sobreveniencia* de lo mental sobre lo físico. Tal sobreveniencia implicaría, según Davidson[25], que si dos eventos difieren en sus propiedades mentales, entonces difieren en sus propiedades físicas. Las propiedades mentales serían causalmente relevantes porque tener o no tener ciertas propiedades mentales, o tener o no tener propiedades mentales en general, implicaría *cambios* en las propiedades físicas, y estos cambios en las propiedades físicas *sí* son causalmente relevantes. Son varios los problemas que le han sido planteados a Davidson en este punto. Señalaremos sólo dos.

23. Fodor (1987 y 1989). Véanse también, por ejemplo, Dretske (1989) y Follesdal (1985).
24. Véase especialmente LePore y Loewer (1987 y 1989), Horgan (1989) y McLaughlin (1989).
25. Véase Davidson (1970, 1985 y 1993).

En primer lugar[26], el que las propiedades mentales impliquen diferencias físicas que, a su vez, pueden implicar diferencias causales es una cosa. Y otra muy distinta que esas diferencias físicas lleguen a implicar las diferencias causales *relevantes* como para que, por ejemplo, algunos eventos mentales puedan considerarse causa de una acción. Se requiere una eficacia causal *más específica* que la ofrecida por Davidson. En segundo lugar, la relación de sobreveniencia que propone Davidson es terriblemente *débil*. Es una simple afirmación de hecho que no tiene *ninguna fuerza modal* capaz de prestar apoyo a contrafácticos de ningún tipo. Como repetidamente ha indicado Kim[27], la sobreveniencia de Davidson sería compatible con la hipotética *eliminación o redistribución* de todas las propiedades mentales. Podríamos imaginar un mundo en el que las piedras pensaran y los humanos fueran como piedras y, mientras se cumpliera que dos piedras tienen los mismos pensamientos en ese mundo sólo si comparten todas sus propiedades físicas, y otras cosas similares, seguiría manteniéndose una sobreveniencia de lo mental sobre lo físico en el sentido de Davidson. Parece necesaria alguna fuerza modal capaz de impedir estos casos. La eficacia causal de lo mental en el monismo anómalo de Davidson está, como vemos, llena de problemas.

3. *Reduccionismos fisicalistas*

Todos los reduccionismos fisicalistas que *no* sean *eliminativistas*, que no rechacen simplemente la realidad de lo mental, tienen una respuesta muy directa al problema de su eficacia causal: los fenómenos mentales son causalmente eficaces simplemente porque *son* fenómenos físicos. No existiría *mayor* eficacia causal posible. Ahora bien, la manera de entender esa respuesta depende de la manera como se entienda tal *reducción* o *identidad* entre lo mental y lo físico. Tendríamos aquí, básicamente, tres alternativas:

1) *Reducciones o identificaciones definicionales:* Afirmarían que toda propiedad mental puede redefinirse con la ayuda de alguna propiedad física. Simbólicamente[28] $(M)(EF)(M=_{Def}F)$. Esta posición, o alguna versión más o menos mitigada de ella, fue mantenida por el conductismo, tanto lógico como psicológico, y por el operacionalismo[29]. Los conceptos

26. Véase Kim (1993a y 1993b) y Sosa (1993).

27. Por ejemplo, en Kim (1978, 1979, 1984b, 1984c, 1989a, 1993a y 1993b).

28. La variable M representa propiedades mentales, la variable F representa propiedades físicas, la variable x representa objetos que pueden ejemplificar las anteriores propiedades, (...) cuantifica universalmente y (E...) lo hace existencialmente; utilizaremos también otros símbolos lógicos habituales.

29. Tres magníficas compilaciones (acompañadas de abundante bibliografía) de trabajos relativos a la mayoría de las posiciones en filosofía de la mente y ciencias cognitivas que estamos mencionando son Block (1980), Lycan (1990) y Rosenthal (1991).

con los que nos referimos a las propiedades que se reducen o identifican definicionalmente se convierten en conceptos sinónimos, y las respectivas propiedades pasan a ser coextensivas. Se pretende que a través de ciertos conceptos físicos se capte, sin residuo alguno de significado, lo expresado a través de los conceptos con los que normalmente nos referimos a las propiedades mentales. Justificar esta pretensión involucraría tanto análisis conceptuales como investigaciones empíricas. Y el peso relativo de cada uno de estos elementos indicaría el carácter más o menos filosófico o empírico de las definiciones ofrecidas.

2) *Reducciones o identificaciones «type*[30]*-type» (o de propiedades)*: Afirmarían que toda propiedad mental puede hacerse coextensiva con alguna propiedad física. Simbólicamente, *(M)(EF)(x)(MxFx)*. Ésta fue, y sigue siendo, la reducción preferida por el fisicalismo. Sobre todo, respecto a los estados mentales no intencionales (sensaciones, etc.)[31]. La fuerza modal y el alcance contrafáctico de estas reducciones puede ser mayor o menor, según la coextensionalidad pueda mantenerse en campos más o menos significativos o amplios de situaciones posibles distintas de las actuales. Habitualmente se piensa que estas reducciones deberían mantenerse en todas las situaciones físicamente posibles y, así, tener el rango que el fisicalismo clásico reservaba a las leyes reductivas[32]. Una teoría T sería reducible a otra teoría T' syss T es lógicamente derivable a partir de T'. Las leyes reductivas conectarían los términos de T (la teoría reducida) con los de T' (la teoría reductora), haciendo posible esa derivación. Desechadas las reducciones o identificaciones definicionales, los enunciados bicondicionales de las reducciones o identificaciones *type-type* ofrecerían aquí una conexión perfecta.

3) *Reducciones o identificaciones* token-token *(o de casos):* Afirmarían que todo aquello que pueda ejemplificar una propiedad mental ejemplifica cierta propiedad física tal que, al ejemplificar esta propiedad física, ejemplifique también aquella propiedad mental. Simbólicamente, *(M)(x)(EF)(MxFx)*. También aquí, la fuerza modal y el alcance contrafáctico de estas afirmaciones podría ser mayor o menor. Nótese que, aunque 1 implique a 2, y 2 implique a 3, las conversas no se cumplen. En

30. Tanto aquí como más abajo, mantenemos la terminología inglesa técnica usual en estos temas. Un *type* es una entidad abstracta, una clase de cosas; un *token* es un particular, una cosa concreta. Extensionalmente, las propiedades son clases de cosas (*types*), y las cosas concretas (*tokens*) ejemplifican propiedades al pertenecer a ciertas clases de cosas (a ciertos *types*).

31. Véase, por ejemplo, Feigl (1967) y Armstrong (1968). Muchas posiciones fisicalistas en un sentido más amplio (posiciones materialistas, podríamos decir) también adoptan este enfoque; véase, por ejemplo, Bunge (1977 y 1980).

32. *Bridge principles*, *bridge laws* o *correlation laws*. Véase al respecto el modelo clásico de reducción de teorías propuesto por Nagen (1961).

particular, las reducciones o identificaciones *token-token* no implicarían ninguna reducción o identificación *type-type*. Podría ocurrir, por ejemplo, que dos objetos o y o' ejemplificaran una misma propiedad mental Mi ejemplificando, respectivamente, propiedades físicas diferentes Fj y Fk, y no existiendo, sin embargo, ninguna propiedad física relevante compartida por ambos objetos (una propiedad como, por ejemplo, FjvFk). El monismo anómalo, el funcionalismo y otras muchas perspectivas en filosofía de la mente sólo admitirían estas reducciones. Y las considerarían suficientes.

Sabemos que todos los intentos (conductistas, operacionalistas, etc.) por conseguir reducciones o identificaciones definicionales de las propiedades mentales han *fracasado*. Ya hemos visto también cómo el monismo anómalo de Davidson necesitaba *complementar* su identificación de eventos mentales con eventos físicos, su peculiar reducción o identificación *token-token*, con algo más a fin de escapar de la acusación de epifenomenalismo. Y que aun eso era *poco*. ¿Conseguirían las reducciones o identificaciones *type-type* asegurar la eficacia causal de lo mental? Por una parte, parece que lo que queremos es la eficacia causal de lo mental *en cuanto mental* y que, por consiguiente, una reducción de las propiedades mentales a ciertas propiedades físicas no haría más que reducir la eficacia causal de lo mental a la eficacia causal de lo físico. Pero, por otra parte, si las propiedades mentales resultaran últimamente identificables con ciertas propiedades físicas, seguramente sería con propiedades físicas con unas características *tan especiales* que no nos resultaría ya nada *extraña* esa identificación. En este sentido, podemos decir que no existiría *mayor* eficacia causal para lo mental que la que se obtendría reduciendo o identificando las propiedades mentales con ciertas propiedades físicas.

De todas formas, de hecho, tampoco *disponemos* actualmente de reducciones o identificaciones fisicalistas efectivas de ninguna propiedad mental. Todos los candidatos son tremendamente circunstanciales, su fuerza modal es bastante restringida y, entre otras cosas, están llenos de cláusulas *caeteris paribus* ineliminables. En cuanto a la reducción de teorías, por otro lado, lo que hemos dicho resulta demasiado idealista y restrictivo. Aparte de que tal vez no sean necesarios los bicondicionales de las reducciones o identificaciones *type-type*, y basten unos simples condicionales, la exigencia de que la teoría reducida ha de poder *derivarse lógicamente*, con ayuda de las leyes reductivas, de la teoría reductora es excesiva. Lo que suele derivarse, en todos los casos paradigmáticos de reducción de teorías, es una *versión modificada y corregida* de la teoría reducida [33]. La reducción de teorías se convierte, con todo esto, en

33. Este punto se encuentra perfectamente desarrollado, en directa relación con el problema que nos ocupa, en Churchland, Paul (1989a).

cierto tipo de *explicación* de la teoría reducida en términos de la teoría reductora más que en un asunto de simple derivabilidad lógica. Y es a través de esa explicación como conseguimos la unificación, sistematización y simplicidad ontológica que buscamos en la reducción de teorías.

No sólo no podemos, pues, ofrecer actualmente claras reducciones o identificaciones *type-type* de ninguna propiedad mental, sino que tampoco tendríamos por qué tenerlas aunque hubiéramos conseguido reducir toda la psicología natural a algún conjunto de teorías más básicas. Y, sin embargo, como decíamos antes, tal vez fueran estas identificaciones las que *mejor* aseguraran la eficacia causal de lo mental.

4. *La alternativa de la sobreveniencia*

Ya hemos hablado de cierta relación de sobreveniencia al comentar el monismo anómalo de Davidson. En estos últimos años, la noción de sobreveniencia [34] ha sido filosóficamente muy explotada para caracterizar relaciones no causales de *dependencia* o *determinación* entre propiedades de diverso tipo. Por ejemplo, entre propiedades normativas o evaluativas (de tipo ético, estético, epistemológico, etc.) y ciertas propiedades no normativas ni evaluativas. Generalmente se supone que las relaciones de sobreveniencia deberían ser reflexivas, antisimétricas y transitivas. Si unas propiedades sobrevienen sobre otras, tener las segundas determina el que se tengan las primeras y estas últimas dependen de aquéllas. La distinta fuerza modal de todas estas relaciones, y la distinta forma de entender las mismas propiedades que pueden aquí intervenir, hace que se mantenga una u otra versión conceptual de la noción de sobreveniencia.

Muchos conceptos de sobreveniencia no tienen implicaciones reductivas *type-type*, por ejemplo el de Davidson. Algunos de ellos ni siquiera implicarían reducciones o identidades *token-token*. Esto ocurre, por ejemplo, con conceptos muy *globales* de sobreveniencia en los que se afirma, simplemente, que mundos indiscernibles respecto a las propiedades que constituyan la base de una sobreveniencia deben ser también indiscernibles respecto a las propiedades sobrevenientes. Tampoco nos comprometeríamos con ninguna reducción o identidad *token-token* si definimos nuestros conceptos de sobreveniencia de forma que, indepen-

34. Podríamos emplear en castellano tanto «sobreveniencia» como «superveniencia» (con todos sus derivados) para traducir el término inglés «supervenience» (y sus derivados). Aunque el sustantivo «sobreveniencia» no aparezca explícitamente en nuestros diccionarios (véase, por ejemplo, el *Diccionario ideológico de la lengua española* de Julio Casares), la existencia actual en castellano de abundantes usos ordinarios y jurídicos de términos más relacionados con «sobreveniencia» que con «superveniencia», y la mayor riqueza semántica que ofrecen esos primeros términos (véase el mismo diccionario antes citado) hace que me incline a preferirlos sobre los segundos. Acerca de la noción general de sobreveniencia, de sus múltiples conceptualizaciones concretas y usos, las siguientes referencias serían fundamentales: Horgan (1984 y 1993) y Kim (1984a y 1990a). Véase también Liz (1993). La casi totalidad de los trabajos de Kim sobre este tema se encuentran recientemente recogidos en Kim (1993c).

dientemente de su mayor o menos globalidad, las propiedades sobrevenientes sean ejemplificadas en dominios de individuos *distintos* de los dominios en los que se ejemplifiquen las propiedades de la base de esas sobreveniencias[35]. Todos estos conceptos de sobreveniencia normalmente intentan sustentar posturas filosóficas generales *anti-reduccionistas*, sin romper completamente con el fisicalismo. Pero otros conceptos, en cambio, como a continuación veremos, sí tienen implicaciones reductivas. De cualquier forma, las relaciones de sobreveniencia ofrecen una variedad *no causal* de determinación o dependencia. Y esto las hace interesantes. Y las distingue, por ejemplo, de las ambiguas relaciones de dependencia o determinación causal del epifenomenalismo y del emergentismo.

El problema con el concepto de sobreveniencia de Davidson consistía en que resultaba bastante *débil*. No aseguraba, por ejemplo, que alguien como cualquiera de nosotros, incluso exactamente igual molécula a molécula, tuviera *necesariamente* nuestros mismos pensamientos. Parece que se necesita una conexión más estrecha entre lo mental y lo físico. Asumiendo que los conjuntos de propiedades deben estar cerrados respecto a las operaciones booleanas clásicas, Kim nos propone el siguiente concepto de *sobreveniencia en sentido fuerte*: un conjunto de propiedades A sobreviene fuertemente sobre un conjunto de propiedades B syss, necesariamente, para cualquier x y cualquier propiedad Ai de A, si x tiene Ai, entonces existe una propiedad Bj en B tal que x tiene Bj y tal que, necesariamente, si cualquier otro y tiene Bj, entonces también tiene Ai.

Este concepto de sobreveniencia implicaría que, si las propiedades mentales sobrevienen sobre ciertas propiedades físicas, entonces si un sujeto tiene una propiedad mental M, ha de tener también una propiedad física F tal que, necesariamente (en cualquier situación o circunstancia), si tiene F, tiene también M. A partir de aquí, Kim propone entender las *relaciones causales en las que parecen intervenir fenómenos mentales*[36] en estrecha conexión con ciertas relaciones causales físicas *subyacentes* según el siguiente esquema (M representa una propiedad mental; C, por ejemplo, una acción; F y G propiedades físicas): Ma causa Cb syss 1) {M} sobreviene fuertemente sobre {F}, 2) {C} sobreviene fuertemente sobre {G}, y 3) Fa causa físicamente Gb.

Para que nuestras explicaciones causales consigan captar eficacias causales y relaciones causales genuinas, han de existir *relaciones causales físicas subyacentes* con las que se conecten, al menos, a través de relaciones fuertes de sobreveniencia. Y tales eficacias y relaciones causales resultarán problemáticas en la medida en que esas relaciones de sobreveniencia dejen de ser tan fuertes. Las relaciones aparentemente causales

35. Respecto a estas posibilidades (sobreveniencia fuerte, débil, global, para dominios coordinados, etc.), véase Kim (1984a).

36. Y, más en general, toda relación causal macrofísica. A veces, Kim las llama a todas ellas «causalidades epifenoménicas». Véase Kim (1984b y 1984c).

que creemos descubrir entre las imágenes de un espejo que refleje cierto proceso o entre los síntomas de una enfermedad, etc., no serían así relaciones causales genuinas. Y tampoco lo serían las relaciones entre los fenotipos que señalábamos en el segundo apartado (véase 14 y 15). Sin embargo, tal vez así se entienda mejor cómo en el ejemplo de la explosión de ese mismo apartado (véase 16) sí teníamos tanto una eficacia causal como una relevancia explicativa. Y muchos casos de explicaciones causales que mencionan fenómenos mentales como causa o efecto de algo (muchos casos de los tipos A, B o C) *sí* podrían detectar genuinas relaciones causales. La eficacia causal de las propiedades mentales queda reivindicada, en el planteamiento de Kim, a través de la eficacia causal de ciertas propiedades físicas subyacentes.

Pero si con el concepto de sobreveniencia de Davidson teníamos un problema de debilidad, ahora tenemos uno de *excesiva fortaleza*. Del concepto de Kim de sobreveniencia en sentido fuerte también se desprende que debe existir una propiedad física F# tal que necesariamente sea tenida siempre que se tenga la propiedad M. F# sería la *disyunción* de todas aquellas propiedades físicas que pueden servir como base de una sobreveniencia de M. Lo que obtenemos, generalizando, es algo como (M)(EF#)(x)(MXF#), justamente una *reducción o identidad type-type* (o de propiedades). Aplicando esto al análisis que acabamos de hacer respecto a la eficacia causal de lo mental, resultaría la siguiente equivalencia: M es causalmente eficaz, en general, para la producción de algo que es C syss F# es causalmente eficaz, en un estricto sentido *físico*, respecto a la producción de algo que es G#. La eficacia causal de las propiedades mentales acaba siendo *reducible* (idéntica, coextensiva con) la estricta eficacia causal física de ciertas propiedades físicas. Ahora bien, *¿es esto realmente un problema*?

Podemos hacer aquí las mismas matizaciones que hicimos en relación al reduccionismo fisicalista basado en reducciones o identidades *type-type*. Pese a nuestras primeras intuiciones y prejuicios anti-reduccionistas, pese a no disponer actualmente de ninguna incondicionada reducción efectiva respecto a ninguna propiedad mental y pese a que las propias reducciones interteóricas puedan incluso llevarse a cabo sin tales reducciones, tal vez no exista *mayor eficacia causal para lo mental* que la que podría obtener resultando ser idéntico a algo físico. ¿Por qué, después de todo, esa identidad última tendría que restar eficacia causal a lo mental *en cuanto mental*, en lugar de hacerlo a lo físico *en cuanto físico*?

Hemos examinado brevemente algunas posibles opciones a la hora de armonizar la clausura causal del mundo físico con la eficacia causal de lo mental. Hay otras opciones que no hemos analizado. No obstante, creo que nos hemos fijado en las más importantes.

IV. LA PECULIAR REALIDAD DE LO MENTAL: SOBREDETERMINACIÓN, VIRTUALIDAD, EXTERNALISMO Y REALIZABILIDAD MÚLTIPLE

Pensemos ahora en el enunciado 10. Esta explicación sugiere *dos* nuevos e inquietantes problemas. El primero de ellos tiene que ver con el tipo de relación causal que aquí surge. Aunque no supusiéramos que ayer se acostaron tarde, de todas formas, el que sepamos que siempre son impuntuales habría igualmente causado que imagináramos su retraso en la cita. Y, aunque no supiéramos nada sobre su impuntualidad, el suponer que ayer se acostaron tarde habría igualmente causado el mismo efecto. Claramente nos encontramos ante un caso de *sobredeterminación explicativa y causal*: dos explicaciones y dos causas parecen ser suficientes para la explicación y producción de un mismo efecto. Y no estamos frente a ningún ejemplo extraño o rebuscado. Las sobredeterminaciones explicativas y causales son el pan de cada día por lo que se refiere a nuestra vida mental. Nada similar encontramos en el mundo *físico* ni en las explicaciones causales que solemos hacer respecto a lo que ocurre en él.

En relación al mundo físico, los casos de sobredeterminación no abundan. Y ello sugiere la adopción de cierto *principio de exclusión explicativa y causal* para lo físico según el cual, salvo ciertos casos claros de sobredeterminación, nada puede tener varias explicaciones o causas completas e independientes[37]. Suele indicarse que este principio pondría en serio peligro la relevancia explicativa de lo mental. Si existieran explicaciones causales alternativas que únicamente mencionaran fenómenos no mentales, lo mental no podría ser explicativamente relevante. Y que pone también en serio peligro la eficacia causal de lo mental cuando se asume que siempre deben existir procesos causales de tipo no mental, y en último término físicos, subyaciendo a los procesos causales que involucran fenómenos mentales.

De todas formas, y con independencia de los problemas planteados por lo físico, el anterior principio de exclusión explicativa y causal introduciría serios problemas *dentro* de la misma esfera de lo mental. Lo mental estaría muchas veces sobredeterminado explicativa y causalmente en relación ahora a *sí mismo*. ¿Por qué nuestra vida mental escapa tantas veces al principio de exclusión? ¿No será porque gran parte de las explicaciones que mencionan fenómenos mentales no son realmente explicaciones causales, sino sólo *racionalizaciones*?

El segundo problema se refiere al *tipo de realidad* que habrían de

37. Véase Kim (1990b). Aunque Kim, en general, sólo hable de exclusión explicativa, sus argumentos también podrían apoyar un principio de exclusión causal. Y esto sería conveniente; sobre todo, teniendo en cuenta 1) que la clausura causal de lo físico sería perfectamente compatible con una masiva e implausible sobredeterminación causal y 2) que puede llegar a ser metodológicamente provechoso mantener, durante cierto tiempo, explicaciones alternativas completas e independientes, con lo cual el principio de exclusión causal resultaría ser incluso más básico y fundamental que el de exclusión explicativa.

tener los fenómenos mentales a los que apelamos en estas explicaciones. Sigamos pensando en el enunciado 10. Saber que siempre son impuntuales no parece ser un estado que debamos tener *constantemente presente* en nuestra mente mientras imaginamos que llegarán tarde. Tal vez ni siquiera llegue a estar nunca *presente de la misma forma* como lo está el suponer que ayer se acostaron tarde (por no compararlo con el tipo de presencia que parecen requerir estados como sentir un fuerte dolor, véase 4). Saber que siempre son impuntuales puede, perfectamente, tener el mismo tipo de *inconstante* presencia en la mente que la presencia que tienen estados mentales como saber que Platón escribió «La república» o saber que 45+2=47. Algunos fenómenos mentales tienen una realidad enormemente *virtual*. Y tal vez esté aquí *otra* de las claves de su gran sobredeterminación explicativa y causal.

Pero un enunciado como 10 aún sugiere otros problemas. Saber que siempre son impuntuales parece *implicar* que realmente lo sean. *Saber* que son impuntuales no es lo mismo que simplemente *creer* que lo son. Si realmente no son impuntuales, saber que son impuntuales no puede ser aquí un estado causalmente eficaz simplemente porque *no existe* ningún estado tal. Este es un importante sentido en el que algunos fenómenos mentales presentan rasgos *relacionales o externalistas*. Pero hay otros sentidos. Consideremos, por ejemplo, un enunciado como 3. Que haya un árbol enfrente se propone como condición suficiente para creer que hay un árbol enfrente. En ciertas circunstancias, esto puede ser verdad. Puede incluso servir de base a ciertos principios epistemológicos interesantes. Ahora bien, si nuestro sujeto nunca hubiera tenido árboles enfrente ni nunca hubiera interactuado adecuadamente con árboles reales, de una u otra forma, parece que resultaría muy difícil aceptar que pueda tener creencias sobre los árboles. En esas condiciones, la atribución de creencias sobre los árboles sería sólo eso, una *mera atribución*. Algún tipo de interacción en el pasado, más o menos directa, con árboles reales parece una *condición necesaria* para poder tener creencias sobre ellos. La misma existencia de creencias sobre los árboles, creencias capaces de tener una eficacia causal, dependería así esencialmente de factores relacionales externos a los sujetos.

Esta característica que presentan muchos fenómenos mentales ha sido enfatizada recientemente por numerosos autores, generalmente haciendo uso de *experimentos mentales* del tipo desarrollado inicialmente por Hilary Putnam y Tyler Burge [38]. Todos estos experimentos mentales presentan situaciones contrafácticas, más o menos plausibles empíricamente [39], en las que variaciones en la historia, el entorno o el contexto

38. Véase Burge (1979, 1986, 1989 y 1993) y Putnam (1975). En relación a este tema, véanse también las compilaciones de trabajos de Pettit y McDowell (1986) y de Woodfield (1983).

39. A veces la acción se desarrolla en una «Tierra Gemela», otras veces en sitios tan familiares

lingüístico de cierto sujeto hacen que cambien nuestras consideraciones acerca de cuál puede ser el contenido semántico de sus estados mentales intencionales. Dos sujetos que compartieran el *mismo tipo de estados internos* (físico-químicos, funcionales, etc.) podrían, curiosamente, *no pensar lo mismo*.

Para mantener su eficacia causal, decíamos antes, un estado como saber que siempre son impuntuales no parece requerir la misma actualidad o presencia real que otros estados. Y no sólo esto. Como vemos ahora, su misma existencia, y con ella su eficacia causal, depende fuertemente de factores externos a lo que pueda ocurrir dentro de los sujetos. En muchas ocasiones, lo mental parece tener una realidad tremendamente *virtual y externa a los propios sujetos*.

Pero aquí no acaban los problemas ontológicos. Un *mismo* fenómeno mental, un mismo estado o proceso de tipo intencional o no intencional, parece poder ser tenido por *sistemas de tipo muy diferente*. Tan diferente que acaso no sea posible considerarlos pertenecientes a las mismas clases naturales salvo por lo que se refiera a esos fenómenos mentales que todos ellos manifiestan.

Respecto a muchos fenómenos mentales, lo importante es que se respete cierta *organización funcional*, cierta estructura, pasando a un segundo plano cuáles sean las clases concretas de cosas que son capaces de tener esa organización funcional o esa estructura. Se suele expresar esta idea diciendo que gran parte de los fenómenos mentales son *fenómenos funcionales múltiplemente realizables* en realidades físicas del tipo más diverso. Y la generalización de esta idea es el germen de la perspectiva *funcionalista* en filosofía de la mente y, en general, en todas las ciencias cognitivas [40].

El funcionalismo resulta indudablemente muy *liberal*, pero no es fácil encontrar razones de peso, razones no «chauvinistas», para *restringir* esa liberalidad y negar que una misma creencia, deseo, etc. (incluso, por ejemplo, la creencia y el deseo citados en 6) no puedan ser estados tenidos tanto por un humano como por un sofisticado robot o por un extraterrestre con una composición orgánica totalmente distinta a la nuestra [41].

Estamos diciendo que, aunque los fenómenos mentales siempre sean ejemplificados por realidades físicas, lo cual significa al menos una reducción o identidad *token-token*, puede que esas realidades físicas no

como otro subgrupo lingüístico. Putnam, por ejemplo, es mucho más aficionado que Burge a los viajes interplanetarios del primer tipo, lo cual tal vez no haya favorecido demasiado la correcta comprensión de los problemas.

40. Entre los creadores del funcionalismo, en sus muchas vertientes, deberíamos citar a Fodor (1981), Lewis (1972 y 1980), Putnam (1960, 1965 y 1967) y Sellars (1956).

41. Desde el clásico trabajo de Block (1978), este problema (ser demasiado liberales o caer en un «chauvinismo") sigue siendo muy controvertido. Véase, por ejemplo, Stich (1990).

tengan *nada* relevante en común aparte de ser físicas y de ejemplificar esos fenómenos mentales. Pero aquí surge un problema grave. La realizabilidad múltiple de lo mental impide la existencia de reducciones o identidades *type-type* entre propiedades mentales y propiedades físicas. Y esto pone en serio *peligro* la relevancia explicativa y eficacia causal de lo mental. Si, por ejemplo, 6 y 7 aluden a estos tipos de fenómenos, el enunciado legaliforme 13 se nos convierte en una ley de un tipo muy particular. No sólo ya por incluir una cláusula *caeteris paribus*, sino por mencionar fenómenos cuya realidad se *entrecruza*, en una peligrosa indeterminación, con la realidad física.

Una manera de enfrentarse a este problema puede consistir en *rechazar la realizabilidad múltiple de lo mental*, permitiendo que la simple disyunción de las propiedades físicas relevantes de todas aquellas cosas capaces de ejemplificar una determinada propiedad mental también cuente como una propiedad física genuina. Una segunda alternativa sería aceptar plenamente la realizabilidad múltiple de lo mental y *ampliar el concepto de causalidad* a fin de dar cabida a explicaciones y relaciones causales establecidas a través de propiedades irreductiblemente funcionales. Y una última alternativa, mezcla de las otras dos, consistiría en *asumir esa realizabilidad múltiple resistiéndose, al mismo tiempo, a abrir la puerta de la causalidad a las propiedades funcionales*. En este caso, las *propias* explicaciones y relaciones pretendidamente causales que mencionan fenómenos mentales serían *múltiplemente realizables*. Debería haber diferentes explicaciones y diferentes relaciones causales para cada tipo peculiar, tipo descriptible en términos físicos, de realidades físicas capaces de ejemplificar los fenómenos mentales en cuestión. Una eficacia causal y una relevancia explicativa *distinta*, por tanto, para *cada clase física relevante de sistemas físicos capaces de ejemplificar cierto fenómeno mental*. Las tres opciones anteriores son tremendamente actuales. Pero todas ellas están llenas de problemas. La última parece la más razonable. Sin embargo, con ella la ilusión de que lo mental pueda ser siempre, por sí mismo, explicativamente relevante y causalmente eficaz, *con independencia de las peculiaridades físicas* de los sistemas concretos en los que se encuentre ejemplificado, se desvanece.

V. CIERTA RELEVANCIA EXPLICATIVA *CAETERIS PARIBUS* Y EFICACIA CAUSAL DE LO MENTAL PESE A TODO

Pero ¿a qué tipo de realidad nos estamos refiriendo a través de todos estos matices? Desde luego, es una realidad bastante *distinta* de la realidad que descubrimos en el mundo físico. Algunos autores extienden estos rasgos a todos los fenómenos mentales obteniendo conclusiones *instrumentalistas o eliminativistas*. Lo mental no será más que el resultado

de aplicar una compleja red conceptual, el resultado de interpretar y racionalizar de cierta forma el comportamiento de algunos sistemas. Y a pesar de que esa red conceptual, nuestra psicología natural o *folk psychology*, pueda tener cierto valor instrumental, sería erróneo tomarla teóricamente en serio. El instrumentalismo de Daniel Dennett, el eliminativismo neurofisiológico de los Churchland o el eliminativismo computacional de Stephen Stich serían ejemplos perfectos de esto. Gran parte de la actual corriente conexionista en las ciencias cognitivas también explota profusamente este rasgo de lo mental con propósitos instrumentalistas o eliminativistas (o, en el mejor de los casos, con propósitos ambiguos)[42]. Otros autores, en cambio, no se dejan impresionar tan fácilmente y, *asumiendo esos rasgos como componentes esenciales de lo mental*, intentan nuevas vueltas de tuerca de manera que, aun así, lo mental pueda seguir teniendo *cierta relevancia explicativa y eficacia causal* cuando intervenga por el lado de las *causas* o cuando intervenga por el lado de los *efectos*. Después de todo, ya sabíamos que lo mental era algo *muy especial*. Este era uno de los caminos que quedaron abiertos al discutir la posición de Davidson.

Analizaremos brevemente dos intentos de este tipo, el de Jerry Fodor y el de Fred Dretske. Para ambos, la relevancia explicativa y la eficacia de lo mental, como causa o como efecto causal, debe manifiestarse inevitablemente a través de *adecuadas explicaciones y leyes causales caeteris paribus*. Tanto Fodor como Dretske dirigen sus planteamientos especialmente a esos fenómenos mentales que hemos llamado intencionales. Se asume, y tal vez sin razón, que los otros, los cualitativos o fenoménicos, tienen menos problemas y están mucho más cercanos a lo físico que aquellos. El tratamiento que dan al papel que los estados mentales intencionales pueden desempeñar como efecto causal sería hasta cierto punto *similar* en los dos. No así el papel que asignan a los estados y procesos mentales intencionales como causa; aunque, como señalaremos, sus posiciones podrían llegar a ser también aquí *compatibles*. Comencemos con esto último.

1. *Fenómenos mentales intencionales como causa de la acción y de nuestros pensamientos: los caminos encontrados de Fodor y de Dretske*

Fodor mantiene una *teoría representacional y computacional de la mente*[43]. Según tal teoría, los estados mentales intencionales son rela-

42. Véanse Dennett (1978 y 1989), Churchland, Patricia (1986), Churchland, Paul (1979, 1989a y 1989b) y Stich (1985 y 1990). Respecto al aspecto del conexionismo que acabamos de mencionar, véase Clark (1989). En relación al eliminativismo neurofisiológico de los Churchland, uno de los reduccionismos fisicalistas de tipo eliminativista más interesantes de los últimos años, véase Gomila (1990).

43. Fodor (1975, 1981 y 1987) son obras que ofrecerían una vista panorámica bastante completa de su teoría.

ciones computacionales que un sujeto mantiene con ciertas representaciones y los procesos mentales son operaciones computacionales realizadas con esas representaciones. Las representaciones forman un sistema innato de símbolos con una estructura lingüística, el llamado *lenguaje del pensamiento*, y las relaciones y operaciones computacionales sólo tienen en cuenta las propiedades *sintácticas* de las representaciones, no su semántica. El objetivo de la psicología científica es desentrañar las características sintácticas de ese lenguaje en el que nuestras neuronas procesan la información y hacer explícitas las relaciones y operaciones computacionales que allí se producen. Y puede hacerlo sin conocer apenas nada sobre el cerebro. Lo único que debe tener en cuenta son los *papeles causales* de las representaciones. La psicología es una *ciencia especial* no reducible a otras ciencias más básicas.

Todo esto compromete con una especie de *solipsismo metodológico* que sitúa a la propia base física del sujeto y a las relaciones del sujeto con todo lo que sea externo a él fuera del alcance de la psicología. También obliga a distinguir dos tipos de contenido mental, uno en *sentido amplio* y otro en *sentido restringido*. Los contenidos mentales en sentido amplio son los contenidos semánticos adscritos en la psicología natural, son contenidos tremendamente sensibles a la historia de los sujetos, a sus entornos y a las comunidades lingüísticas a las que pertenezcan. Son los contenidos que se originan cuando *contextualizamos* la vida mental de un sujeto. Los contenidos mentales en sentido restringido surgen de las peculiaridades sintácticas de las representaciones mentales y hacen que, en cada contexto particular, cada representación mental tenga los concretos valores semánticos que tiene. Estas peculiaridades sintácticas se encuentran *materializadas* de alguna forma en nuestros cerebros. Nuestros cerebros llevan a cabo los *papeles causales* que las definen. Y la eficacia causal de los contenidos mentales se debe a estos papeles causales.

El computacionalismo de Fodor trata de llevar hasta sus últimas consecuencias la *analogía* entre la mente y los computadores clásicos. La mente es un tipo muy especial de computador. La mente es un computador con su *software* y su *hardware*. El *software* de nuestra mente es tremendamente complicado[44], su *hardware* es orgánico. Los computadores adquieren, almacenan y procesan información de una manera indiscutiblemente física, pero muchos de sus estados son también susceptibles de ser descritos en términos de un contenido semántico. En mi computador, por ejemplo, ahora mismo se está procesando y almacenando esta frase. Los procesos computacionales ofrecen una curiosa forma de *conciliar la causalidad con la semántica* (y, si convertimos a

44. A veces, por cierto, tan complicado que sería imposible su mismo análisis científico. Respecto a esto, concretamente a la imposibilidad de aclarar completamente la estructura y dinámica de los sistemas psicológicos centrales de fijación de creencias, véase Fodor (1983).

nuestro computador en un robot, a ambas con la *acción*) a través de una adecuada *sintaxis*.

Así es como, en el planteamiento de Fodor, se abriría sitio la *eficacia causal de lo mental* a través de ciertas relaciones causales físicas que se desarrollen en el sujeto. Y la *relevancia explicativa de lo mental* también tendría su sitio. Las explicaciones y leyes *caeteris paribus* de la psicología natural son una perfecta guía para la psicología científica. Bastaría con «restringir» adecuadamente los contenidos mentales que son aludidos en ellas para obtener buena parte de la estructura computacional que ha de tener el sistema interno de representaciones que hemos llamado el *lenguaje del pensamiento*.

Dretske sigue un camino distinto del de Fodor. Los estados mentales intencionales son, para él, causalmente eficaces al actuar como *causas estructurantes* de ciertos procesos en virtud de sus peculiares contenidos semánticos en sentido amplio[45]. La distinción entre causas estructurantes y *causas desencadenantes* resulta aquí fundamental. Hay dos maneras como podemos hablar de la causa de cierto proceso consistente en que una causa C cause un efecto E. Si nos preguntamos por lo que hace que ese proceso ocurra en cierto momento determinado, nos estaríamos preguntando por sus causas desencadenantes, por lo que ha causado la producción de un E. Pero si nos preguntamos por lo que hace que C cause E, lo que buscamos son las causas estructurantes del proceso; es decir, las causas de la propia conexión causal entre C y E. Las causas desencadenantes más próximas de nuestras acciones y pensamientos son estudiadas por la *neurofisiología*. Ahora bien, la pregunta: ¿por qué esas causas desencadenantes consiguen tener los efectos causales que tienen?, no puede ser respondida *sólo* por la neurofisiología. La respuesta a esta pregunta estaría determinada por la *función* que desempeñen esas causas desencadenantes en la organización estructural a la que pertenezcan los procesos consistentes en que esas causas tengan esos efectos. Cuando esa causa sea un estado mental intencional, por ejemplo una creencia, la respuesta a esta segunda pregunta estará determinada por el peculiar contenido semántico, por la peculiar función representacional, que tenga tal creencia.

Según Dretske, en la naturaleza hay relaciones *objetivas* que conforman el hecho de que algo tenga un determinado contenido semántico, un significado. Nuestros estados mentales intencionales tienen un contenido semántico fijado de manera natural. Y estos contenidos *estructuran* los procesos causales desarrollados por nuestra acción y nuestros pensamientos. Las *razones* de nuestra acción y de nuestros pensamientos son sus causas estructurantes. En ciertos sistemas, esa estructura se adquiere a través de una historia evolutiva. En otros, mediante un diseño artificial.

45. Véase Dretske (1983, 1988 y 1993).

En nuestro caso, el *aprendizaje individual* tendría una importancia decisiva, incluyendo aquí el aprendizaje de una lengua con la que hablamos acerca de los contenidos semánticos de nuestros propios pensamientos. Lo que creemos, deseamos, etc., es así *causalmente eficaz* respecto a lo que hacemos y a las otras cosas que creemos, deseamos, etc., porque es *la causa de la estructura* que hace posible que ciertos estados internos a nosotros adquieran *control* sobre determinados movimientos de nuestro cuerpo y sobre otros estados internos. Nuestras razones conforman las estructuras causales responsables de nuestras acciones y pensamientos. Y cuando *explicamos* la acción y nuestros pensamientos en base a *razones*, hacemos alusión a esas razones en cuanto sus causas estructurantes.

Dretske, decíamos, elige un camino distinto que el de Fodor para reivindicar la relevancia explicativa y la eficacia de lo mental como causa. Fodor se centra en el *sujeto*. Dretske en el *contexto*, en el entorno del sujeto, su historia y su comunidad lingüística. Sin embargo, ambos caminos podrían, tal vez, llegar a *juntarse* en algún punto. No podemos aquí extendernos más, pero el caso es que si *conectamos* los contenidos en sentido restringido de los que nos hablaba Fodor (esto es, esas características sintácticas de las representaciones mentales capaces de dar lugar, en cada contexto, a unos valores semánticos determinados) con estas estructuras de las que ahora nos habla Dretske (estructuras causales internas al sujeto causadas, a su vez, por el hecho de que ciertos de sus estados tengan el contenido semántico en sentido amplio que tienen), lo que acaso resulte es un *único camino de doble dirección*.

2. *Estados mentales intencionales como efecto causal de algo no mental: de nuevo Fodor y Dretske, esta vez por el mismo camino*

Pero ¿qué hay de lo mental, especialmente de los estados mentales intencionales, como *efecto causal de algo no mental*? Tanto para Fodor como para Dretske hay que situar el *origen* del significado de nuestros lenguajes en el significado o contenido que pueden tener ciertos estados mentales intencionales[46]. Y hay que buscar esos peculiares contenidos semánticos en las *relaciones causales e informacionales* que un sujeto es capaz de establecer con su entorno.

Fodor negaba que los procesos mentales tuvieran acceso a las propiedades semánticas, en sentido amplio, de las representaciones mentales. Pero esto no es negar que existan y se deban intentar aclarar esas propiedades semánticas. Lo que ocurre es que su estudio *no* pertenece al dominio propio de la psicología. Fodor *no* cree posible llegar a establecer

46. Se trataría del clásico programa de Grice (1957) consistente, a grandes rasgos, en reducir la semántica a psicología. Schiffer (1972) es una magnífica presentación y desarrollo de tal programa, y Schiffer (1989) un alegato acerca de la imposibilidad de llevarlo finalmente a cabo.

condiciones naturalistas necesarias y suficientes para que un estado de cosas pueda significar algo. Esas condiciones tendrían que relacionar, por ejemplo, ejemplificaciones del símbolo «agua» con ciertas propiedades de la sustancia H_2O. Pero hasta que no conozcamos todas las cosas del universo, con sus propiedades relevantes, no podrán conocerse con precisión todas esas relaciones. La naturalización completa de la semántica es *prácticamente imposible*. Sin embargo, Fodor admite que puedan establecerse ciertas *condiciones suficientes* que permitirían entender cómo los significados, los contenidos mentales y no mentales, *no escapan* al orden natural del mundo.

Según Fodor[47], el *símbolo* «árbol», por ejemplo, significaría árbol si, entre otras cosas, nada que no sea un árbol puede causar ejemplificaciones de «árbol» a menos que los árboles puedan hacerlo. Esta condición exigiría una *dependencia asimétrica* entre la ley según la cual algunas cosas que no son árboles causan ejemplificaciones de «árbol» y la ley según la cual los árboles causan esas ejemplificaciones. La primera ley no podría ser verdadera si no lo es la segunda (volvemos a insistir en que todas estas relaciones causales y leyes serían relaciones causales y leyes causales *caeteris paribus*). Fodor pretende mostrar así cómo los símbolos y, en particular, los estados mentales intencionales podrían tener un contenido en sentido amplio sumamente *robusto* y *selectivo*. El problema de la robustez consistiría en explicar cómo un símbolo (por ejemplo, «árbol»), significando lo que significa, puede mantener ese significado a pesar de que muchas de sus ejemplificaciones no sean causalmente producidas por cosas que caen bajo su extensión (por ejemplo, por postes eléctricos en la lejanía). Por otro lado, el problema de su carácter selectivo (a veces llamado «problema del error o de la disyunción») consistiría en distinguir casos en los que ciertas ejemplificaciones de un símbolo son erróneamente producidas (por ejemplo, un poste eléctrico en la lejanía que causa ejemplificaciones de «árbol») de casos en los que las ejemplificaciones de un símbolo pueden ser correctamente producidas por varios tipos de cosas, teniendo así el símbolo un significado disyuntivo (por ejemplo, truchas, salmones, lucios, etc., causando ejemplificaciones del símbolo «pez»). La condición impuesta por Fodor sería crucial para solucionar estos problemas. Los símbolos mantendrían robustamente su significado mientras todo aquello que, sin caer bajo su extensión, pueda causar ejemplificaciones suyas (por ejemplo, postes eléctricos en la lejanía causando ejemplificaciones de «árbol») no pueda haberlas causado a menos que esas ejemplificaciones puedan ser causadas por cosas que sí caen bajo su extensión (en nuestro ejemplo, por árboles). Y la selección de un significado excluyente se realiza también en la dirección de esa dependencia. Quedaría excluido del significado de un símbolo todo aquello

47. Fodor (1990). Una interesante crítica al enfoque de Fodor se encuentra en Putnam (1992).

que no pueda causar ejemplificaciones suyas a menos que otras cosas puedan hacerlo (quedarían excluidos así los postes eléctricos en la lejanía del significado del símbolo «árbol», pero no las truchas, salmones, lucios, etc., del significado del símbolo «pez»). Estas consideraciones serían *suficientes*, según Fodor, para entender cómo algunos fenómenos mentales, en particular los robustos y selectivos contenidos semánticos de nuestros estados mentales intencionales, puede ser un *efecto causal de algo no mental*.

Volvamos ahora nuevamente con Dretske. Dretske explica el origen natural del contenido semántico de nuestros pensamientos insistiendo más que Fodor en las *relaciones informacionales*[48]. Las relaciones causales generan flujos de información. Y estos flujos de información hacen de ciertos estados de cosas *signos naturales* de otras cosas a las que representan. A partir de aquí, ciertos sistemas procesadores de información consiguen que en algunos de sus estados internos se genere la *función de representar* algo. Estas funciones representacionales son independientes de cualquier convención. Son fijadas por la manera como los signos naturales son *usados* por el sistema del cual son parte. Un estado de cosas adquiere la función de representar algo cuando el hecho de representarlo selectivamente cobra importancia para el propio desarrollo y buen funcionamiento del sistema.

Dretske se enfrenta aquí a los mismos problemas que veíamos en Fodor. El contenido semántico es *robusto* y muy *selectivo*. Y se sirve de la distinción establecida entre representar y tener la función de representar[49] para explicar estos rasgos. Una cosa es *representar algo* y otra tener la *función de representarlo*. No se puede representar *erróneamente* algo a menos que se tenga la función de representarlo. Y la función de representar algo puede ser sumamente *selectiva* y puede mantenerse *robustamente* a pesar de que se represente algo erróneamente. Una vez adquirida esa función, una vez convertida la información en contenido semántico, el *error* se debe simplemente a un *mal funcionamiento* del sistema.

Sería difícil resumir las múltiples polémicas surgidas en torno a las perspectivas de Fodor y Dretske. Más aún cuando sus trabajos en estos temas se encuentran aún en fase de desarrollo. De todas formas, insistamos una vez más sólo en un punto. La casi totalidad de las explicaciones, leyes y relaciones causales sobre las que se apoyan ambos planteamientos son de tipo *caeteris paribus*. Sea cual sea la perspectiva que adoptemos en estos temas, parece incuestionable que las *interpretaciones* y las *consideraciones normativas*, dependientes de cierta concepción de lo que cabe

48. Véase especialmente Dretske (1981).
49. La evolución de sus ideas en este punto pueden rastrearse a través de Dretske (1981, 1986 y 1988).

o no esperar una vez que se conocen las circunstancias concretas en las que se desarrolla una vida mental, resultan también aquí decisivas. Lo mismo que en las explicaciones cotidianas que llevamos a cabo en el marco de la psicología natural[50].

VI. CONCLUSIONES

Como hemos visto, las explicaciones en las que mencionamos fenómenos mentales pueden ser de muy variado tipo. En ellas, los fenómenos mentales pueden llegar a ser mencionados como causa, como efecto o como causa y efecto. Distinguíamos así tres tipos de explicaciones causales que mencionan fenómenos mentales, nuestros tipos A, B y C. Los fenómenos mentales pueden ser estados mentales, intencionales o no intencionales, o procesos mentales. Y en nuestras explicaciones también podemos encontrarnos con fenómenos no mentales, con estados o procesos más o menos cercanos a lo físico.

Las explicaciones causales que aluden a fenómenos mentales tienen ciertamente un gran parecido formal al resto de nuestras explicaciones causales. Admiten algunas formulaciones bastante objetivas y contrastables, prestan apoyo a predicciones y son subsumibles en generalizaciones legaliformes. Pero se trata siempre de explicaciones sometidas a cláusulas *caeteris paribus*. Hemos señalado que para que puedan ser auténticas explicaciones causales, y puedan llegar a detectar genuinas relaciones causales, 1) han de existir realmente procesos causales subyacentes, y 2) el nexo entre su *explanans* y su *explanandum* no puede ser meramente conceptual. Pero, aparte de lo espinoso que pueda ser determinar estas cosas, existen otros problemas que no dejan ver claro cómo esas explicaciones, y las leyes asociadas a ellas, pueden ser causales en el mismo sentido en el que lo son las explicaciones y leyes causales desarrolladas respecto al mundo físico.

El mundo físico parece estar causalmente cerrado y regulado por cierto principio de exclusión explicativa y causal. Lo mental, por otro lado, evita persistentemente todos nuestros intentos reductivos y presenta numerosos rasgos que lo apartan de lo físico. Rasgos como la sobredeterminación, la virtualidad, el externalismo y la múltiple realizabilidad.

50. Llegamos al final de nuestro trabajo y, también, al final de las notas. De todas las referencias bibliográficas que hemos ido presentando, habría varias de una importancia fundamental: Burge (1993), Churchland, Paul (1989b), Fodor (1987, 1989 y 1990), Davidson (1980 y 1993), Dretske (1981, 1988 y 1993), Horgan (1989), Kim (1984a, 1984b, 1984c, 1990b, 1993a, 1993b y 1993c), LePore y Loewer (1987), LePore (1989), Searle (1992) y Sosa (1984 y 1993). Algunas de estas referencias se encuentran en el reciente libro editado por Heil y Mele (1993) que lleva por título, justamente, *Mental Causation*. En su conjunto, este recomendable libro es una perfecta muestra del estado actual de la cuestión en este terreno.

Hemos examinado algunos planteamientos filosóficos generales que intentan preservar la relevancia explicativa y la eficacia causal de lo mental a través de esa estricta disciplina impuesta por lo físico. Hemos hablado del epifenomenalismo, del emergentismo, del monismo anómalo de Davidson, de distintas variedades de reduccionismos no eliminativistas y de la alternativa de la sobreveniencia. Y también hemos mencionado la posibilidad de mantener posturas instrumentalistas y eliminativistas. Finalmente, nos hemos detenido a presentar los enfoques de Fodor y Dretske. Independientemente del planteamiento filosófico general que se mantenga respecto a las relaciones de lo mental con lo físico, estos últimos enfoques asumían todos los anteriores rasgos característicos de lo mental, consiguiendo cierta relevancia explicativa y eficacia causal para los fenómenos mentales intencionales cuando intervienen como causa o efecto de algo. Pero aquí esa relevancia explicativa y eficacia causal necesitaba depender siempre de condiciones *caeteris paribus*, con lo que acabábamos volviendo a un sitio muy semejante al punto de partida.

¿Cuál es, después de todo esto, nuestro balance final? Tras la primera impresión de desconcierto, al menos tres cosas deberían estar claras.

La *primera* es que si tuviéramos una teoría naturalizada del contenido, la mayoría de los problemas que hemos discutido se resolverían por sí solos. Los análisis que se hagan de esa clase tan peculiar de explicaciones causales que mencionan fenómenos mentales son más directamente dependientes de nuestras concepciones semánticas que de nuestras concepciones acerca de la causalidad. Sea lo que sea la causalidad, la semántica introduce una infinidad de nuevos problemas. En casi todo nuestro trabajo ha quedado sin determinar con precisión la noción de «relaciones causales genuinas». Pero, como hemos visto, aun así, los problemas se multiplicaban sin cesar, sobre todo respecto a las propiedades semánticas de algunos fenómenos mentales. El caso es que no tenemos actualmente ninguna teoría naturalizada del contenido completamente satisfactoria. Y por ello, además de cargar con los problemas generales de la causalidad (incluyendo aquí el problema mismo del sentido que pueda tener la expresión «relaciones causales genuinas») y con los problemas específicos generados por la existencia de explicaciones causales que se sitúan fuera del estricto mundo físico (como ocurre cuando se mencionan fenómenos mentales), tenemos que enfrentarnos con casi todos los problemas de la semántica.

La *segunda* es que tal vez lo que mejor asegure la eficacia causal de lo mental sea la identificación de las propiedades mentales con ciertas propiedades físicas. Y que no hay razones contundentes para negar que, entonces, lo mental dejaría de ser causalmente eficaz en cuanto mental. Pues, de darse esa identidad, por razones análogas podríamos también decir que lo físico dejaría de ser causalmente eficaz en cuanto físico. La identidad es muy ambiciosa, cierto. Pero se lo puede permitir.

La *tercera* es que, a menos que simplemente nos declaremos instrumentalistas, nuestras valoraciones de la eficacia causal de lo mental tendrán que derivarse de su relevancia explicativa. Y que, a menos que abracemos el eliminativismo, resulta indudable que muchas de las explicaciones y leyes que mencionan fenómenos mentales son enormemente relevantes sin que actualmente dispongamos, ni acaso podamos nunca disponer, de sustitutos aceptables que no mencionen tales fenómenos mentales.

La primera de las anteriores ideas hace comprensible la pluralidad de opciones y planteamientos que encontramos en estos temas. Las otras dos ideas son más sustantivas. Pero actúan en sentidos opuestos. Lo que nos promete la segunda nos lo quita la tercera, y viceversa. Tenemos así una *posible ontología reduccionista* y una *práctica explicativa anti-reduccionista*. Pero sólo quien siempre sienta la urgente necesidad de elegir se sentirá apurado por esta situación.

BIBLIOGRAFÍA

Álvarez, S. (1990), «Contextualidad de la relación causal», en Pérez, 1990.

Armstrong, D. (1968), *A Materialist Theory of the Mind*, Routledge & Kegan Paul, London.

Bennet (1988), *Events and their Names*, Hackett Publishing Company, Indianapolis.

Block, N. (1978), «Troubles with Functionalism», en Rosenthal, 1991.

Block, N. (1980) (ed.), *Readings in Philosophy of Psychology*, Harvard University Press, Cambridge.

Bogdan, R. (1986) (ed.), *Belief*, OUP, Oxford.

Broad, C. (1925), *The Mind and its Place in Nature*, Routledge and Kegan Paul, London.

Broncano, F. (1993), «What on Earth Would Be Lost if Truth and Rationality Were Lost?», *Conferencia SOFIA-93*, La Laguna.

Bunge, M. (1977), «Emergence and the mind»: *Neuroscience*, 2.

Bunge, M. (1980), *The Mind-Body Problem*, Pergamon, Oxford. V.e.: *El problema mente-cuerpo*, Tecnos, Madrid, 1985.

Burge, T. (1979), «Individualism and the mental»: *Midwest Studies in Philosophy*, 4.

Burge, T. (1986), «Individualism and Psychology»: *Philosophical Review*, 95.

Burge, T. (1993), «Mind-Body Causation and Explanation», en Heil y Mele, 1993.

Churchland, P. (1986), *Neurophilosophy. Toward a Unified Science of the Mind/Brain*, MIT Press, Cambridge.

Churchland, P. (1979), *Scientific Realism and the plasticity of Mind*, Cambridge University Press, Cambridge.

Churchland, P. (1989a), *A Neurocomputational Perspective. The Nature of Mind and the Nature of Science*, MIT Press, Cambridge.

Churchland, P. (1989b), «Folk Psychology and the Explanation of Human Behavior»: *Philosophical Perspectives*, *3*.
Clarke, A. (1989), *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing*, MIT Press, Cambridge.
Fodor, J. (1975), *The Language of Thought*, Harper & Row, New York. V. e.: *El lenguaje del pensamiento*, Cátedra, Madrid, 1984.
Fodor, J. (1981), *RePresentations*, MIT Press, Cambridge.
Fodor, J. (1983), *The Modularity of Mind*, MIT Press, Cambridge. V. e.: *La modularidad de la mente*, Morata, Madrid, 1986.
Fodor, J. (1987), *Psychosemantics*, MIT Press, Cambridge.
Fodor, J. (1989), «Making Mind Matter More»: *Philosophical Topics*, *17*.
Fodor, J. (1990), *A Theory of Content*, MIT Press, Cambridge.
Follesdal, D. (1985), «Causation and Explanation: A Problem in Davidson View on Action and Mind», en LePore y McLaughlin, 1985.
Davidson, D. (1963), «Actions, Reasons, and Causes», en Davidson, 1980.
Davidson, D. (1970), «Mental Events», en Davidson, 1980.
Davidson, D. (1973), «The Material Mind», en Davidson, 1980.
Davidson, D. (1974), «Psychology as Philosophy», en Davidson, 1980.
Davidson, D. (1980), *Actions and Events*, OUP, Oxford.
Davidson, D. (1993), «Thinking Causes», en Heil y Mele, 1993.
Dennett, D. (1978), *Brainstorms*, MIT Press, Cambridge.
Dennett, D. (1987), *The Intentional Stance*, MIT Press, Cambridge.
Drestke, F. (1981), *Knowledge and the Flow of Information*, MIT Press, Cambridge. V. e.: *Conocimiento e información*, Salvat, Barcelona, 1987.
Drestke, F. (1983), «Reasons and Causes»: *Philosophical Perspectives*, *3*.
Drestke, F. (1986), «Misrepresentation», en Bogdan, 1986.
Drestke, F. (1988), *Explaining Behavior. Reasons in a World of Causes*, MIT Press, Cambridge.
Drestke, F. (1993), «Mental Events as Structuring Causes of Behaviour», en Heil y Mele, 1993.
Esquivel, J. (1982) (ed.), *La polémica del materialismo*, Tecnos, Madrid.
Ezquerro, J. (1984), «Algunas razones para naturalizar la razón», en Pérez, 1984.
Gomila, A. (1990), «El materialismo eliminativista de los Churchland»: *Contextos*, *VIII/15-16*.
Grice, H. (1957), «Meaning»: *Philosophical Review*, *66*.
Feigl, H. (1967), *The «Mental» and the «Physical»*, University of Minnesota Press, Minneapolis.
Heil, J. y A. Mele (1993) (eds.), *Mental Causation*, Clarendon Press, Oxford.
Honderich, T. (1982), «The argument for Anomalous Monism»: *Analysis*, *42*.
Horgan, T. (1984) (ed.), *The Concept of Supervenience in Contemporary Philosophy*, Spindel Conference Supplement, *Southern Journal of Philosophy*, *22*.
Horgan, T. (1989), «Mental Quausation»: *Philosophical Perspectives*, *3*.
Horgan, T. (1993), «From Supervenience to Superdupervenience: Meeting the Demands of a Material World»: *Mind*, Otoño.
Huxley, T. (1898), *Method and Results. Collected Essays,* volume I, Macmillan, London.

Johnston, M. (1985), «Why Having a Mind Matters», en LePore y McLaughlin, 1985.
Kim, J. (1978), «Supervenience and Nomological Incommensurables»: *American Philosophical Quarterly, vol. 15, 2.*
Kim, J. (1979), «Causality, Identity, and Supervenience in the Mind-Body Problem»: *Midwest Studies in Philosophy, 4.*
Kim, J. (1984a), «Concepts of Supervenience»: *Philosophy and Phenomenological Research, vol. XLV, 2.*
Kim, J. (1984b), «Supervenience and Supervenient Causation»: *Southern Journal of Philosophy, 22.*
Kim, J. (1984c), «Epiphenomenal and Supervenient Causation»: *Midwest Studies in Philosophy, IX.*
Kim, J. (1987), «"Strong" and "Global" Supervenience Revisited»: *Philosophy and Phenomenological Research, vol. XLVIII, 2.*
Kim, J. (1988), «Supervenience for Multiple Domains»: *Philosophical Topics, vol. XVI, 1.*
Kim, J. (1989a), «The myth of Nonreductive Materialism»: *Presidential Address delivered before the eighty-seventh Annual Central Division Meeting of the American Philosophical Association*, Chicago, 1989.
Kim, J. (1989b), «Mechanism, Purpose, and Explanatory Exclusion»: *Philosophical Perspectives, 3.*
Kim, J. (1990a), «Supervenience as a Philosophical Concept»: *Metaphilosophy, vol. 22, 1-2.*
Kim, J. (1990b), «Explanatory Exclusion and the Problem of Mental Causation», en Villanueva, 1990.
Kim, J. (1993a), «Can Supervenience and "Non-strict Laws" Save Anomalous Monism?», en Heil y Mele, 1993.
Kim, J. (1993b), «The Non-reductivist Troubles with Mental Causation», en Heil y Mele, 1993.
Kim, J. (1993c), *Supervenience and Mind*, Cambridge Univ. Press, Cambridge.
LePore, E. y McLaughin, B. (1985) (eds.), *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, Basil Blackwell, Oxford.
LePore, E. y Loewer, B. (1987), «Mind Matter»: *The Journal of Philosophy, 93.*
LePore, E. y Loewer, B. (1989), «More on Making Mind Matter»: *Philosophical Topics, vol. XVII, 1.*
Lewis, D. (1972), «Psychophysical and Theoretical Identifications»: *Australasian Journal of Philosophy, 50.*
Lewis, D. (1980), «Mad Pain and Martian Pain», en Block, 1980.
Liz M. (1993), *La vida mental de algunos trozos de materia. Teorías de la sobreveniencia de lo mental*, Laertes, Barcelona.
Lombard, L. (1986), *Events. A metaphysical Study*, Routledge & Kegan Paul, London.
Lycan, W. (1990) (ed.), *Mind and Cognition. A Reader*, Basil Blackwell, Cambridge.
McLaughlin, B. (1989), «type Epiphenomenalism, Type Dualism, and the Causal Priority of the Physical»: *Philosophical Perspectives, 3.*
Morgan, C. (1923), *Emergent Evolution*, William & Norgate, London.
Mosterin, J. (1978), *Racionalidad y acción humana*, Alianza, Madrid.

Nagel, E. (1961), *The Structure of Science*, Harcourt, New York.
Pérez, J. (1990) (ed.), *Conocimiento y acción*, Servicio de Publicaciones de la Universidad de Salamanca, Salamanca.
Pettit, P. y J. McDowell (1986)(eds.), *Subject, Thought and Context*, Clarendon Press, Oxford.
Popper, K. y J. Eccles (1977), *The Self and its Brain*, Springer-Verlag, Berlín. V. e.: *El yo y su cerebro*, Labor, Barcelona, 1982.
Putnam, H. (1960), «Minds and Machines», en Putnam, 1975b.
Putnam, H. (1965), «Brains and Behavior», en Putnam, 1975b.
Putnam, H. (1967), «The Nature of Mental States», en Putnam, 1975b.
Putnam, H. (1975a), «The Meaning of "Meaning"», en Putnam, 1975b.
Putnam, H. (1975b), *Mind, Language and Reality. Philosophical Papers,* volume 2, Cambridge University Press, Cambridge.
Putnam, H. (1992), *Renewing Philosophy*, Harvard University Press, Cambridge.
Quesada, D. (1984), «Las afirmaciones ontológicas y la psicología de las actitudes proposicionales»: *Teorema*, *vol. XIV/3-4.*
Quintanilla, M. (1989), *Tecnología: un enfoque filosófico*, Fundesco, Madrid.
Rosenthal (1991)(ed.), *The Nature of Mind*, Hutchinson, London.
Schiffer, S. (1972), *Meaning*, Oxford University Press, Oxford.
Schiffer, S. (1984), *Minds, Brains and Science: The 1984 Reith Lectures*, Harvard University Press, Cambridge. V. e.: *Mentes, cerebros y ciencia*, Cátedra, Madrid, 1985.
Schiffer, S. (1987), *Remnants of Meaning*, MIT Press, Cambridge.
Schiffer, S. (1992), *The Rediscovery of the Mind*, MIT Press, Cambridge.
Searle, J. (1980), «Minds, Brains and Programs»: *The Behavioral and Brain Sciences*, *3*.
Sellars, W. (1956), «Empirism and the Philosophy of Mind», en Sellars, 1963.
Schiffer, S. (1963), *Science, Perception and Reality*, Routeledge and Kegan Paul, London. V. e.: *Ciencia, percepción y realidad*, Tecnos, Madrid, 1971.
Sosa, E. (1984), «Mind-Body Interaction and Supervenient Causation»: *Midwest Studies in Philosophy*, *IX*.
Sosa, E. (1993), «Davidson's Thinking Causes», en Heil y Mele, 1993.
Stich, S. (1985), *From Folk Psychology to Cognitive Science*, MIT Press, Cambridge.
Stich, S. (1990), *The Fragmentation of Reason*, MIT Press, Cambridge.
Stoutland, F. (1985), «Davidson on Intentional Behaviour», en LePore y McLaughlin, 1985.
Tuomela, R. (1989), «Collective Action, Supervenience, and Constitution»: *Synthese*, *80*.
Vázquez, M. y Liz, M. (1993), *Más allá de lo natural y de lo artificial* (ms., pendiente de publicación).
Villanueva, E. (1990)(ed.), *Information, Semantics and Epistemology*, Basil Blackwell, Oxford.
Woodfield, A. (1982)(ed.), *Thought and Object*, Clarendon Press, Oxford.

ELIMINATIVISMO Y EL FUTURO DE LA PSICOLOGÍA POPULAR

Josefa Toribio Mateas

I. INTRODUCCIÓN

Es asombrosa la cantidad y variedad de acciones que pueden desencadenar ciertos estados mentales. Tómese como ejemplo el miedo y véase sin más lo que le ocurre a las tías en ciertos cuentos:

> Por más que hagamos, tía tiene miedo de caerse de espaldas; y su inocente manía nos afecta a todos, empezando por mi padre que fraternalmente la acompaña a cualquier parte y va mirando el piso para que tía pueda caminar sin preocupaciones, mientras mi madre se esmera en barrer el patio varias veces al día, mis hermanas recogen las pelotas de tenis con que se divierten inocentemente en la terraza, y mis primos borran toda huella imputable a los perros, gatos, tortugas y gallinas que proliferan en casa ... (Cortázar, 1970, 39).

El *miedo* de la tía a caerse de espaldas causa la movilización de todos los miembros de la familia. El *deseo* de ayudar en esta agonía les lleva a realizar diferentes acciones encaminadas a evitar accidentes que pudieran provocar la caída. Si tuviéramos que explicar por qué el padre inspecciona el piso incansablemente, tendríamos que decir que la causa es que a) *sospecha* de la existencia de objetos que pueden propiciar el accidente; b) *desea* que su hermana no tenga que preocuparse por ellos, y c) *confía* en que su compañía hará desaparecer esos temores.

Lo mismo puede decirse de la madre, las hermanas o los primos. Todos ellos *creen* que la limpieza y la ausencia de obstáculos contribuirá a evitar el desastre y todos ellos *desean* evitarlo. Basándonos en esas creencias y deseos, podemos predecir fácilmente su conducta. La tía ca-

minará despacio y angustiadamente, con el deseo permanente de no encontrar situaciones amenazantes. Las acciones de los demás, llevados por el convencimiento de que algo terrible puede ocurrir si la tía finalmente tropezara y cayera, serán un puro ir y venir por la casa en la búsqueda del orden perfecto.

Aunque muy probablemente el contenido de los miedos, creencias, sospechas y deseos sea distinto, nuestra vida está llena de explicaciones y predicciones de este tipo. Son explicaciones obvias, de sentido común. Nadie necesita ser un(a) psicólogo/a profesional para explicar, por ejemplo, que Mariano fue a la fiesta porque *deseaba* ver a Dolores y *creyó* que ella se encontraría allí. Y como nadie necesita del diploma en Psicología para explicar y predecir en estos términos, decimos que esta manera de dar cuenta de las acciones propias y ajenas constituye una *Psicología Popular*.

El término «Psicología Popular» denota así un cuerpo de explicaciones y predicciones de la conducta que se caracteriza esencialmente por su apelación a estados mentales con contenido —deseos, creencias, miedos, intenciones, intuiciones y, en general, estados psicológicos construidos en términos proposicionales— como causas de tal conducta. Es una explicación típica de la Psicología Popular la de que Pedro llegó tarde a su cita con Manuel porque *creía* que habían quedado a una hora diferente. Es tal creencia la que, de acuerdo con la Psicología Popular, constituye la causa interna de la conducta de nuestro amigo.

Cuando yo predigo que Lola abrirá la nevera sobre la base de su *deseo* de comer un pedazo de tarta y su *creencia* de que la tarta se encuentra en la nevera, también utilizo una herramienta predictiva típica de la Psicología Popular. El deseo de Lola, en las condiciones descritas, nos permite predecir su acción y hacerlo tomando tal deseo como el elemento causal de la misma.

Para el eliminativista, sin embargo, cuando Pedro explica su retraso aduciendo que él *creía* que la cita era a las nueve, o nosotros predecimos la conducta de Lola en términos de su *deseo* de comer tarta, no está/estamos haciendo sino utilizar una herramienta conceptual absolutamente vacía de contenido. No existe ninguna entidad física que pueda hacerse corresponder con la creencia «tener una cita a las nueve» o con el deseo «comer un trozo de tarta». Nada que no sea físico puede actuar como causa. Luego la apelación a tales estados intencionales como causalmente responsables de nuestra conducta es una apelación improcedente.

El significado de *materialismo eliminativo* o, lo que es lo mismo, *eliminativismo* aparece espléndidamente recogido en las siguientes palabras de Paul Churchland:

> El materialismo eliminativo es la tesis de que nuestra concepción de sentido común acerca de los fenómenos psicológicos constituye una teoría radicalmente falsa, una

teoría tan fundamentalmente defectuosa que tanto los principios como la ontología de esa teoría serán eventualmente reemplazados, en lugar de ligeramente reducidos, por una neurociencia completa (Churchland, P. M., 1981, 67).

La tesis se sitúa en la misma línea que los discursos clásicos de Feyerabend (1963), Quine (1966) o Rorty (1965). Fue, de hecho, Feyerabend quien abrió este provocativo debate arguyendo que las categorías mentales que aparecen en las explicaciones de sentido común son términos vacíos, i.e., no refieren a ninguna realidad física, y que la adscripción de estados mentales caracterizados intencionalmente es, por tanto, un procedimiento erróneo de explicación de la conducta. El eliminativismo en su versión contemporánea —y la tesis de Paul Churchland en particular— acentúa además el carácter *teorético* de esas explicaciones intencionales.

La Psicología Popular es una teoría, se arguye. Cuando yo doy cuenta del hecho de que Ana hace ejercicio diciendo que Ana desea llevar una vida sana, mi explicación funciona como una instancia particular de una generalización que tiene un carácter nomológico. Esas generalizaciones tienen como dominio estados mentales con un contenido semántico determinado, *i.e.*, actitudes proposicionales. La generalización que sustenta mi explicación de la conducta de Ana es del tipo siguiente: para cualquier individuo x, cualquier deseo p y cualquier creencia q, si x (Ana) desea p (llevar una vida sana) y x cree que si q, entonces p (Ana cree que si hace ejercicio, lleva una vida sana), entonces —en ausencia de otros deseos o estrategias que entren en conflicto con ello— x (Ana) realiza q (hace ejercicio). Son generalizaciones de este tipo las que forman el entramado teórico de la Psicología Popular. Y son estas generalizaciones las que son consideradas falsas por el/la eliminativista [1].

Las razones para mantener una postura tan agresivamente anti-intuitiva son complejas y no siempre —nunca, dirán algunos— convincentes. Los argumentos que la sustentan y que amenazan así el futuro de la Psicología Popular pueden agruparse, no obstante, en cuatro categorías diferentes. En primer lugar están aquéllos basados en la supuesta *incapacidad* explicativa de la Psicología Popular. Paul Churchland (1981) ha sido el más vigoroso defensor de esta línea argumentativa (sección II).

En segundo lugar, y también con Paul Churchland a la cabeza, tenemos la, más importante, línea reduccionista. Esta segunda línea aboga por la eliminación de las entidades que pueblan la ontología de la Psicología Popular sobre la base de que tales entidades —deseos, creencias y, en general, las actitudes proposicionales— son irreducibles a entidades que puedan ser consideradas neurocientíficamente aceptables (sección III).

La tercera clase de argumentos gira en torno a la noción de organi-

1. La caracterización de la Psicología Popular como una teoría sigue siendo una idea central incluso en los últimos escritos de Paul Churchland (cf. Churchland, 1991).

zación causal del sistema cognitivo humano e intenta mostrar que tal organización es completamente diferente a la postulada por la Psicología Popular. Stephen Stich es el principal representante de esta postura (Stich, 1978, 1982, 1983) que tiene a su base una cierta teoría sobre la individuación y atribución de creencias (sección IV).

La prolongación computacional de estas ideas constituye nuestra cuarta y última categoría de argumentos (Sección V). La reformulación computacional de la tesis es la siguiente. Los modelos computacionales denominados conexionistas aparentemente corroboran la tesis eliminativista en el sentido de que, en ellos, las relaciones *input-output* pueden explicarse sin referencia a ninguna categoría intencional. Si la clase de mecanismos a la que tendría que remitirse el proyecto de modelado computacional de los procesos cognitivos está mejor representada por estos modelos, *i.e.*, si el conexionismo es correcto como modelo cognitivo, entonces las categorías intencionales de la Psicología Popular han de ser eliminadas del marco explicativo de una Psicología científica.

El análisis y desarrollo de estos argumentos, así como de diferentes réplicas a los mismos —que aparecen como subsecciones dentro de cada apartado— nos ayudará a perfilar con más detalle la postura eliminativista y a entender mejor los futuros riesgos o venturas de la Psicología Popular.

No obstante, antes de entrar en los detalles es necesaria una precisión muy importante. Está relacionada con los límites precisos de aquello que parece ser criticado por los diferentes argumentos eliminativistas. No es siempre fácil dilucidar si el objeto a eliminar es la Psicología Popular en su acepción más vaga y amplia —explicaciones causales de acciones que apelan a creencias y deseos— o el modelo sentencial de representación mental que parece subyacer a ese tipo de explicaciones. Este último es el modelo sentencial reivindicado por Fodor en términos de su famosa hipótesis del Lenguaje del Pensamiento (Fodor, 1975).

El pilar básico de esta hipótesis lo constituye la caracterización de los procesos cognitivos como procesos computacionales. Los procesos computacionales se definen, a su vez, en términos de representaciones. Una representación es un tipo de configuración física muy especial, una configuración física que tiene una lectura sintáctica y una lectura semántica.

En el marco del paradigma clásico, que es en el que Fodor desarrolla su hipótesis, esta imagen de la mente como un computador supone una interpretación de los estados intencionales en términos de estados que involucran símbolos de un lenguaje mental —*mentalés*—. De ahí el calificativo de *modelo sentencial*.

De acuerdo con el modelo sentencial, mi creencia de que hay fresas con nata en la nevera conlleva mi estar en algún tipo de relación computacional con el símbolo de *mentalés* correspondiente a «hay fresas con nata en la nevera». El contenido de tal estado intencional es el con-

tenido de ese símbolo en *mentalés*, y el hecho de que sea una creencia —en lugar de un deseo o una duda— está determinado por la naturaleza de la relación computacional con el resto de mis estados mentales y/o mi conducta. Los símbolos de este lenguaje mental están dotados de una sintaxis combinatoria del tipo de la que rige en el cálculo proposicional y son implementados físicamente por patrones de excitación celular. De esta manera, los procesos que subyacen a las relaciones entre estados intencionales, se arguye, son, en última instancia, procesos físicos.

La primera categoría de argumentos y, en particular, los argumentos esgrimidos por Paul Churchland en los primeros días del embate eliminativista, parecen ser claros ataques a la Psicología Popular en su versión de simple sentido común. Los demás ataques son menos claros en lo referente a su objetivo y, para muchos de ellos, éste no parece ser otro que el más refinado modelo sentencial de representación mental que acabamos de exponer. Este es el caso de Patricia Churchland (1981, 1986), militante del eliminativismo en tanto en cuanto la verdad de la Psicología Popular dependa de la existencia de un lenguaje del pensamiento a la Fodor.

Los argumentos de Patricia Churchland, asimilados con frecuencia, erróneamente, a los de Paul Churchland, son una defensa de la tesis de que el modo central de representación es no sentencial. Como tales, no constituyen un apoyo explícito a la tesis eliminativista, pero contribuyen a mostrar que el modelo sentencial de representación mental choca frontalmente con la investigación empírica en las diferentes neurociencias y con los modelos computacionales que esas investigaciones neurológicas favorecen. En ese sentido, tales argumentos pertenecen indirecta pero claramente a nuestro marco taxonómico.

II. INCAPACIDAD EXPLICATIVA: PAUL CHURCHLAND

Dos son las acusaciones principales que Paul Churchland lanza contra la Psicología Popular en esta nuestra primera categoría de argumentos:

a) la existencia de importantes fallos explicativos y predictivos, y

b) el estancamiento empírico, *i.e.*, el hecho de que las explicaciones típicas de la Psicología Popular sigan siendo hoy las mismas que fueron hacen dos mil años. Esta mezcla de incapacidad e inmovilidad explicativa hace de la Psicología Popular, según Paul Churchland, un programa de investigación degenerativo cuya eliminación estaría justificada.

Centrémonos en el primero de los aspectos mencionados y utilicemos la descripción del problema que el mismo Paul Churchland nos proporciona:

> Como ejemplos de fenómenos mentales centrales e importantes que siguen siendo parcial o totalmente un misterio en el marco de la Psicología Popular, considérese la

> naturaleza y dinámica de las enfermedades mentales, la facultad de imaginación creativa ... la naturaleza y funciones psicológicas del sueño ... la construcción interna de una imagen visual tridimensional ... la rica variedad de ilusiones perceptuales ... el milagro de la memoria ... la naturaleza del proceso mismo de aprendizaje ... (Churchland, P. M., 1981, 73).

En relación con estos y otros fenómenos, la Psicología Popular carece completamente de poder explicativo. Esta razón por sí sola no parece suficiente para justificar que la Psicología Popular es una teoría *falsa*, pero muestra, al menos, que es una teoría superficial y profundamente limitada. Cualquier otra teoría que presentase los mismos defectos habría sido abandonada. Es sólo porque la Psicología Popular forma esa parte tan importante de nuestra vida cotidiana en el trato con los demás que estamos dispuestos a seguir defendiéndola por encima de sus claras limitaciones teóricas.

Ante ese ataque de Paul Churchland, dos han sido las respuestas principales. La primera intenta mostrar que la apelación a fenómenos del tipo de los que menciona Paul Churchland como fenómenos que la Psicología Popular debería explicar, es una apelación ilegítima. De acuerdo con Terence Horgan y James Woodward (Horgan y Woodward, 1985), el rango de fenómenos que una teoría ha de explicar viene generado por sus propios recursos teóricos. En el caso de la Psicología Popular, los recursos teóricos son extremadamente simples y están orientados básicamente a la explicación de conductas *comunes* en términos de creencias y deseos. Son conductas comunes aquellas que se desarrollan en el marco de la vida cotidiana y no requieren de un profesional de la Psicología para su interpretación o explicación. Que Rosario fue al cine Excelsior *porque* quería ver *Danzón*, o las visitas de Manolo a su madre motivadas por el *deseo* de degustar su sabrosísima leche frita, son ambas conductas comunes en el sentido que nos interesa aquí. La introducción en el dominio explicativo de la Psicología Popular de procesos mucho más complicados no está en modo alguno justificada y, por tanto, tampoco está justificada una crítica basada en la incapacidad para dar cuenta de ellos[2].

La segunda línea argumentativa utilizada por Horgan y Woodward para contrarrestar la acusación de incapacidad explicativa es la siguiente. Si bien la Psicología Popular arroja escasa luz sobre los problemas de imaginación, imágenes e ilusiones visuales, etc., al menos proporciona ciertas herramientas conceptuales que son utilizadas por la Psicología Cognitiva para desarrollar teorías más refinadas que pueden dar cuenta de los mismos. Así, aunque el esclarecimiento de las relacio-

2. De hecho, algunos autores han cuestionado el propio carácter teórico que Paul Churchland atribuye a la Psicología Popular. Cf. Dennett (1991), Greenwood (1991), Gordon (1986) y Goldman (1989).

nes existentes entre Psicología Popular y Psicología Cognitiva no sea un asunto trivial, siempre es posible defender una tesis moderada según la cual el dominio de la Psicología Cognitiva incluye el dominio de la Psicología Popular en el sentido mencionado, *i.e.*, suministrando herramientas y problemas teóricos. Con esta tesis por bandera, la segunda acusación de Paul Churchland, la acusación de estancamiento de la Psicología Popular pierde también gran parte de su fuerza porque *i*) dada la mayor riqueza teórica de la Psicología Cognitiva, es mucho más difícil demostrar que no es explicativamente plausible, y *ii*) la Psicología Cognitiva ha evolucionado lo suficiente como para estar completamente libre de cualquier acusación de estancamiento y esterilidad (cf. Sterelny, 1990, 149-154).

La apelación a creencias y deseos inconscientes representa, por ejemplo, un progreso empírico evidente en nuestro marco explicativo, una herramienta que permite dar cuenta de complicadas cadenas causales donde antes no era posible (cf. Nisbett y Ross, 1980). La estructura de estas explicaciones, sin embargo, se encuadra perfectamente dentro del modelo de explicación intencional propio de la Psicología Popular.

Las interrelaciones teóricas entre Psicología Popular y Psicología Cognitiva se pueden constatar en los cambios significativos que —como esta apelación al inconsciente— han tenido lugar en la forma de explicar la conducta humana. Estos cambios han sido propiciados por fallos o irregularidades en el modelo explicativo de la Psicología Popular y los efectos de las teorías que han intentado dar cuenta de esos fallos han revertido asimismo sobre ese modelo explicativo, haciéndolo más rico. Dada esta interrelación teórica y dado que la Psicología Cognitiva no está ni ha estado en absoluto estancada, la acusación de estancamiento para con la Psicología Popular se hace mucho más débil.

No obstante, incluso si se rechaza la premisa primera de esta réplica —la que permite ver la Psicología Popular como parte integrante de la Psicología Cognitiva—, existe aún una línea más radical de defensa. De acuerdo con esta línea —de nuevo defendida por Horgan y Woodward— es la misma noción de *progreso* la que no tiene demasiado sentido cuando se aplica la Psicología Popular. Lo que convierte la Psicología Popular en *popular* es el hecho de estar constituida por generalizaciones causales arquetípicas que se aplican una y otra vez para explicar instancias de conductas humanas. La innovación no es una característica de esas generalizaciones, como tampoco lo es de otros juicios causales no psicológicos que hacemos continuamente:

> Tanto nosotros como nuestros antepasados consideramos que el impacto de una piedra causó que la maceta se rompiera, que la carencia de agua causó la muerte del camello ... que el calor causa que el agua hierva, etc. Ninguno de estos juicios forma parte de una teoría empíricamente (rápidamente) progresiva y, sin embargo, pare-

> ce absurdo concluir (atendiendo sólo a esa razón) que son probablemente falsos (Horgan y Woodward, 1985, 402-403).

No parece, pues, que las razones aducidas en esta primera categoría de argumentos sean suficientemente conclusivas para establecer una tesis eliminativista como la reivindicada por Paul Churchland [3]. De hecho, él mismo reconoce que el problema principal planteado por la Psicología Popular es su inconmensurabilidad con el resto de las teorías científicas que tienen al ser humano como objeto explicativo, teorías como la física, la biología evolutiva y, especialmente, las neurociencias. Entramos así en nuestra segunda categoría de argumentos, aquéllos destinados a mostrar la falsedad de la Psicología Popular basándose en la imposibilidad de *reducir* sus categorías teóricas a categorías neurocientíficamente aceptables.

III. LA REDUCCIÓN IMPOSIBLE

1. *Paul Churchland*

No podemos imaginar que un enfoque neurocientífico verdaderamente adecuado de nuestras vidas mentales vaya a proporcionar categorías teóricas que se correspondan detalladamente con las categorías de nuestro sentido común. De acuerdo con esto, lo que debemos esperar es que el sistema más viejo será simplemente eliminado, en lugar de reducido, por una neurociencia completamente desarrollada (Churchland, P. M., 1988, 43).

Una vez que ese desarrollo neurológico haya sido alcanzado, las explicaciones de nuestra conducta dejarán a un lado los términos *deseo* y *creencia* para incorporar términos de diferentes estados neurológicos o de acividades neurales en áreas determinadas (Cf. Churchland, P. M., 1988, 44-45).

A pesar del carácter visionario de estas afirmaciones, Paul Churchland quiere demostrarnos, recurriendo a la historia de la ciencia, cuán lejos están de ser un simple episodio de ciencia ficción. Numerosos casos históricos muestran cómo entidades o concepciones teóricas que fueron aceptadas sin duda alguna durante largos periodos de tiempo, acabaron desapareciendo de nuestro inventario ontológico del mundo. El fluido calórico de los siglos XVIII y XIX desapareció para dejar paso a la cinética molecular. La misma suerte corrió el flogisto ante la llegada de la química contemporánea, por no mencionar brujas, demonios y esferas celestiales. Todos ellos son ejemplos fehacientes de un cambio profundo en los

3. En Baker (1987) y McGinn (1989), especialmente el capítulo dos, pueden encontrarse argumentos adicionales en este sentido.

compromisos ontológicos de diferentes sistemas teóricos. Si estos cambios han tenido lugar en los ámbitos de la física, la química, la astronomía o la medicina, no existe ninguna razón, arguye Paul Churchland, para pensar que un cambio de tal calibre en la Psicología es impensable.

Después de todo, para que un reducción interteórica sea posible en absoluto es necesario:

> ... un conjunto de reglas, «reglas de correspondencia» o «leyes puente» —en terminología standard—, que establezcan una correspondencia entre los términos de la teoría obsoleta (T_o) y un subconjunto de las expresiones de la teoría nueva o reductora (T_n). Estas reglas guian la aplicación de aquellas expresiones seleccionadas en T_n de la siguiente manera: podemos aplicar con toda libertad esas expresiones en todos aquellos casos en donde normalmente aplicaríamos sus dobles en T_o de acuerdo con la regla de correspondencia ...
> En segundo lugar, e igualmente importante, una reducción exitosa tiene idealmente como resultado el que, a través de la proyección de términos efectuada por las reglas de correspondencia, los principios centrales de To (aquéllos de importancia semántica y sistemática) se hacen corresponder asimismo con enunciados generales de Tn que son *teoremas* de T_n (Churchland, P. M., 1979, 81).

Una vez establecidas estas condiciones, resulta evidente que la reducción de la Psicología Popular a una ciencia como la neurología es imposible. Y tal imposibilidad, de acuerdo con Paul Churchland, muestra por sí sola la falsedad de la Psicología Popular. El hecho de que la Psicología Popular es una teoría falsa implica además que las categorías teóricas que la componen están vacías de contenido —pues en caso contrario tal contenido podría ser traducido en el vocabulario neurofisicalista— y, por tanto, que no existen cosas tales como creencias, deseos, miedos, esperanzas o dudas.

El esquema de la argumentación es, pues, como sigue: si la Psicología Popular es verdadera, entonces es reducible a neurociencia. Pero la Psicología Popular no es reducible a neurociencia. Luego la Psicología Popular no es verdadera.

A este caso particular de razonamiento en *modus tollens* se le aplica entonces un particular punto de vista semántico según el cual el significado de un término está determinado por su papel teórico. De acuerdo con este punto de vista, el significado de un término teórico como *creencia* está determinado por el conjunto de leyes que hacen intervenir ese concepto en la explicación de la conducta. Pero si, después de todo, esas leyes no son tales; si la teoría no es verdadera, entonces los términos que componen la teoría se convierten en términos vacíos, desprovistos de referencia alguna. Y si este es el caso, deberíamos abandonar tanto las formas de explicación como la ontología que la Psicología Popular presupone. Hasta aquí el argumento principal desarrollado por Paul Churchland.

2. *Patricia Churchland*

Como dije en la *Introducción*, lo que convierte en especial el caso de Patricia frente a Paul Churchland es el hecho de hacer depender *explícitamente* la verdad de la Psicología Popular de la verdad de la hipótesis fodoriana que postula un *lenguaje del pensamiento*. En este sentido, es un claro ejemplo de lo que Andy Clark ha denominado *eliminativismo sentencial* frente al *eliminativismo intencional* característico de Paul Churchland (cf. Clark, 1993) [4], ya que su objetivo de crítica es precisamente el modelo sentencial. Recuérdese que el modelo sentencial se caracteriza fundamentalmente por postular la existencia de una estructura interna computacional de carácter lingüístico como única forma posible de justificar la conducta de los sistemas intencionales —véase más arriba—.

De forma esquemática, la posición de Pat Churchland es la siguiente: si la mente es una obra de ingeniería lingüística del tipo que la hipótesis fodoriana del lenguaje del pensamiento requiere, entonces la Psicología Popular es una teoría verdadera. Pero el paradigma lingüístico de procesamiento de información es falso, luego la Psicología Popular también lo es. Como vemos, el *modus tollens* es, en este caso, ligeramente diferente en contenido.

Pueden rastrearse al menos dos argumentos en los que Patricia Churchland intenta justificar el carácter no-sentencial de las representaciones mentales. El primero, y más importante, es un argumento de epistemología evolutiva planteado en los siguientes términos. La conducta de muchos animales y ciertamente la de los niños que aún no han aprendido a hablar, parece un ejemplo claro con el que justificar que las propiedades *ser un sistema cognitivo* y *actuar en virtud de representaciones de carácter lingüístico* no son coextensivas. Ante esto, se puede responder aduciendo que, previamente a la adquisición de un lenguaje, o en su ausencia, los procesos cognitivos responsables de esa conducta son no sentenciales, pero que una vez adquirido tal lenguaje —en el caso de los seres humanos adultos— el carácter de los procesos cognitivos cambia esencialmente.

Esta respuesta, argumenta Patricia Churchland, es claramente implausible porque implica la existencia de un salto cualitativo que es absolutamente insostenible desde el punto de vista de la *evolución* de nuestras capacidades lingüísticas. Igualmente insostenible es la respuesta defendida por Fodor. Según Fodor, los procesos cognitivos de niños y de, al menos algunos, animales son *también* procesos sentenciales porque son

4. Incluso Paul Churchland parece haber desradicalizado su postura en los últimos tiempos acercándose más al grupo de eliminativistas sentenciales (cf. Churchland, P. M., 1991). Una buena selección de artículos en los que se discute este y otros muchos aspectos tanto de Patricia como de Paul Churchland es el volumen editado por Robert McCauley, *The Churchlands and Their Critics* (McCauley [ed.], en preparación).

procesos constituidos en términos de un lenguaje del pensamiento independientemente de, o previamente a, la adquisición de un lenguaje público. De hecho, arguye Fodor, el aprendizaje de los conceptos que se expresan en un lenguaje público sólo es posible porque poseemos innatamente estos mimos conceptos formulados en ese lenguaje del pensamiento. O, para ser más específicos, porque poseemos los recursos necesarios para formular, en *mentalés*, las hipótesis que definen la extensión de los conceptos pertenecientes al lenguaje público.

Una postura como ésta, sostiene Pat Churchland, no es más que una reducción al absurdo de la tesis sentencialista. Sostener que nuestro repertorio conceptual incluye *innatamente* el concepto de, por ejemplo, *máquina de escribir* es una posición lo suficientemente absurda como para restar toda plausibilidad a la hipótesis que la sustenta (Cf. Churchland, P. S., 1980 y 1986, 388-389). Y para aquellos que no consideren este tipo de afirmaciones suficientemente auto-refutativas, Pat Churchland vuelve a utilizar consideraciones evolutivas. La atribución de un lenguaje del pensamiento a animales, se defiende, es un claro anacronismo evolutivo porque los procesos cognitivos de tipo sentencial pertenecen claramente a un momento evolutivo posterior:

> El comefrases ha sido ciertamente un recién llegado en el esquema evolutivo de las cosas y debe haberse creado a partir de una organización cognitiva preexistente de carácter no sentencial o, quizá deberíamos decir, debe haber evolucionado a partir de estructuras preadaptativas no sentenciales. Ser un comefrases a todos los niveles, incluso los más básicos, implica o bien que los procesos cognitivos son devoradores de frases desde el final de los tiempos, lo cual es implausible, o que los comefrases no tienen una herencia cognitiva proveniente de especies más primitivas, lo cual es asimismo implausible dada la evolución del cerebro (Churchland, P. S., 1986, 388).

El segundo argumento que mencionábamos intenta mostrar cómo un enfoque sentencial de las representaciones mentales conlleva problemas insolubles tanto a nivel de la adscripción de creencias cuanto al nivel de organización de conocimiento. Desde un punto de vista sentencialista, la adscripción de creencias sólo está justificada si la persona formula esas creencias en *mentalés*, *i.e.*, la noción de creencia se identifica siempre con el pensamiento explícito de tal creencia. La imagen resultante es la de un cuerpo de información almacenada lingüísticamente de tal manera que los enunciados que no pertenecen a ese conjunto no pueden ser considerados como creencias o conocimiento.

De acuerdo con esta tesis, la noción de creencia o conocimiento implícito resulta inexistente porque lo que caracteriza tal conocimiento o tales creencias es precisamente el hecho de que *no se pueden* expresar o formular proposicionalmente. Así, si Pedro no ha concebido jamás la creencia de que las estrellas son alfileres, Pedro no cree tal cosa. Y esta

tesis acerca de las creencias implícitas, denominada *austera* por Patricia Churchland, se corresponde bastante mal con el hecho de que una vez preguntado Pedro si él cree que las estrellas son alfileres, su respuesta es un no inmediato. La facilidad y rapidez con que se producen este tipo de reacciones, sostiene Pat Churchland, muestra que la ausencia de una representación sentencial con respecto a distintos tipos de información no constituye una condición suficiente para restarle su condición de creencia o conocimiento (cf. Churchland, P. S., 1986, 391-392).

El hecho que Pat Churchland subraya, y que juega el papel principal en este argumento, es el hecho de que una gran parte de nuestro conocimiento pertenece al grupo del conocimiento práctico. Es «saber-cómo», no «saber-que», *i.e.*, es un tipo de conocimiento cuya característica definitoria es precisamente que no puede formularse lingüísticamente, que —a diferencia del «saber-que»— no puede enmarcarse en el corsé lingüístico de un conjunto de enunciados y reglas. Una vez diagnosticado el problema en estos términos no es difícil observar la inadecuación de un modelo sentencial de representación [5]. De hecho, defiende Pat Churchland, la aparente insolubilidad de uno de los problemas principales en filosofía e inteligencia artificial relacionado con este tipo de cuestiones, el denominado problema del *marco* (*frame problem*), tiene su origen en esta extensión injustificada del modelo «saber-que» o conocimiento sentencial al conjunto total del conocimiento de un sistema. Pero una vez que el modelo sentencial haya sido eliminado y sustituido por una perspectiva neurofilosófica podremos comprobar hasta qué punto los problemas de creencias y conocimiento tácitos no son más que el fruto de una concepción excesivamente simbólica (cf. Churchland, P. S., 1986, 392-395).

Como dije en un principio, los argumentos de Pat Churchland, aunque de alguna manera menos contundentes que los de Paul, son más cautos en cuanto al objetivo de ataque. Su eliminativismo *sentencial*, sin embargo, conserva la misma tesis de irreductibilidad que el eliminativismo *intencional* de Paul Churchland y puesto que es esta tesis la que ha atraído la mayoría de las críticas, en ella nos concentraremos [6].

3. *Y si no es reducible, ¿qué?*

En el apartado II, cuando presentamos las réplicas a la primera categoría de argumentos, comprobamos que todas ellas intentaban desacreditar las acusaciones que Paul Churchland había lanzado contra la Psicología Popular. Sin embargo, cuando el problema es la irreducibilidad, las ré-

5. Churchland y Dennett, desde presupuestos diferentes coinciden, sin embargo, en el carácter pragmático que rodea el cuerpo explicativo / predictivo de la Psicología Popular. Cf. Dennett (1991).

6. Para una revisión crítica de otros aspectos de la obra de Patricia Churchland pueden consultarse, por ejemplo, von Eckardt (1984) y Kitcher (1984).

plicas tienen un carácter completamente diferente. En este caso se acepta la acusación, *i.e.*, que la Psicología Popular no es reducible a neurociencia —segunda premisa de nuestro *modus tollens*—, pero se intenta resistir la inferencia negando la primera premisa, *i.e.*, negando que la reducción a vocabulario neurocientífico sea una condición necesaria de la verdad de la Psicología Popular. Este es el *espíritu* general de los diferentes contra-argumentos que expondremos a continuación.

Un ejemplo arquetípico de este tipo de respuesta anti-eliminativista es el representado, de nuevo, por T. Horgan y sus colaboradores (cf. Horgan y Woodward, 1985; Graham y Horgan, 1988). Dada la múltiple realizabilidad de los estados mentales y las múltiples diferencias fisiológicas entre individuos, se concede que la reducción de la Psicología Popular a neurociencia —incluso una reducción específica para nuestra especie— es extremadamente improbable. Sin embargo, se arguye, ésta no es razón suficiente para concluir que la Psicología Popular es una teoría falsa. Existen soluciones alternativas que, sin renunciar al componente materialista representado por las explicaciones neurofisiológicas, no son reductivas. Una de estas alternativas es el monismo anómalo de D. Davidson. La tesis monista de Davidson defiende la identidad entre eventos mentales particulares y eventos neurológicos particulares, pero niega la existencia de leyes puente sistemáticas que enlacen tipos de eventos mentales y tipos de eventos neurológicos.

Independientemente de la plausibilidad o implausibilidad de la tesis davidsoniana, el hecho de que, desde un punto de vista conceptual, sea posible mantener un compromiso materialista sin renunciar a las entidades que postula la Psicología Popular muestra que los argumentos eliminativistas basados en la imposibilidad de reducción no son válidos en absoluto.

En un artículo más reciente Horgan y Graham defienden una nueva versión de realismo acerca de la Psicología Popular denominada *fundamentalismo sureño*. En este caso la argumentación no incorpora apelación alguna a posibles tesis materialistas. Se basa, por el contrario, en la noción dennettiana de sistema intencional reformulada en lo que los autores denominan sistema intencional *resonante* (SIR), i.e., un sistema «cuyo completo repertorio conductual es lo suficientemente rico, medio ambientalmente complejo y *prima facie* racional como para que, de acuerdo con criterios epistémicos *ordinarios*, *basados en la conducta*, de atribución de actitudes de la Psicología Popular, no haya ninguna duda sobre si el sistema tiene o no tales actitudes» (Horgan y Graham, 1991, 114).

Para poder ser considerado un sistema que *verdaderamente* sustenta creencias es suficiente con ser un sistema intencional resonante, y la evidencia disponible garantiza el hecho de que los seres humanos son sistemas de ese tipo. Esto significa que, ante la presencia de condiciones de

carácter epistémico que entren en conflicto con la caracterización de los seres humanos como SIR, la conclusión razonable no es la negación de la tesis fundamentalista —y, por tanto, la negación de que existen tales cosas como deseos y creencias—, sino más bien la negación de que tales condiciones sean después de todo un requisito genuino para ser considerado un sistema intencional resonante.

Lo que diferencia este argumento de la línea dennettiana de defensa de la Psicología Popular es el compromiso *realista* de la propuesta. El instrumentalismo de D. Dennett considera que deseos y creencias son estados putativos que pueden ser caracterizados como «ficciones idealizadas» o «constructos lógicos» de la misma manera que lo son los componentes en un paralelogramo de fuerzas (cf. Dennett, 1978, 1981a, 1981b, 1987, 1991). Esta línea hace de la noción de acto racional su concepto central. Así, dada la existencia de sistemas racionales genuinos, como lo son los seres humanos, es necesario postular ciertos estados representacionales que puedan entrar en explicaciones causales de las acciones racionales llevadas a cabo por esos sistemas. Tal reconocimiento no conlleva, de acuerdo con Dennett, ningún compromiso ontológico respecto al tipo de entidades con el que tales estados representacionales han de identificarse. De esta manera Dennett pretende salvar a la Psicología Popular de cualquier riesgo de falsificación que una reducción cientifista pueda acarrear e intenta mostrar que las críticas procedentes de los defensores de tesis reduccionistas no son lo conclusivas que pudieran parecer en un principio.

La respuesta de Paul Churchland a este tipo de consideraciones constituye un argumento general en contra del funcionalismo. Si la forma de reivindicar la Psicología Popular es convertirla en un conjunto de descripciones de los estados y procesos cognitivos que nos permiten predecir correctamente la conducta, pero sin asumir compromiso alguno con respecto a los detalles de su implementación, lo que tenemos es una descripción funcional, demasiado abstracta para ser susceptible a la refutación. Tal inmunidad es ciertamente sospechosa si queremos mantener el carácter empírico de la teoría. La alquimia, arguye Paul Churchland, podría haberse constituido en una teoría inmune a toda refutación proveniente de la química corpuscular si el nivel de descripción adoptado hubiera sido un nivel funcional (cf. Churchland, P. M., 1981, 80).

Ahora bien, una descripción funcionalista de cualquier fenómeno no es necesariamente inmune a la refutación. De hecho, como apunta Kim Sterelny, toda la biología es funcionalista y no por ello es una teoría vacía de contenido o demasiado abstracta para ser refutada por la experiencia (cf. Sterelny, 1990, 152). Una defensa acérrima de los valores de las descripciones funcionalistas cuando la materia de que se trata es la Psicología Popular puede encontrarse asimismo en Jackson y Petit, «In Defence of Folk Psychology» (Jackson y Petit, 1990) y en el primer capítulo de *Psychosemantics* (Fodor, 1988).

Existe así toda una línea argumentativa que insiste en que el poder explicativo y predictivo de la Psicología Popular es absolutamente independiente de los detalles de implementación neurofisiológica. Lo que caracteriza a las explicaciones psicológicas, se arguye, es el hecho de que aportan *unidad conceptual* a eventos y procesos que no podrían ser unificados de otra manera —bajo descripciones a un nivel físico, por ejemplo—, y esto las convierte en necesarias y, por tanto, no eliminables de nuestro repertorio conceptual. De hecho, la unificación explicativo-causal que las descripciones en términos de deseos y creencias aportan a la Psicología Popular cumple el mismo papel que la unificación que supone en física la explicación de ciertos fenómenos a través de conceptos como temperatura o energía (cf. Blackburn, 1991)[7].

No obstante, incluso si, dado el carácter de las explicaciones en la Psicología Popular, existiera un riesgo continuo de inadecuación y error, *i.e.*, incluso si este tipo de argumentos no fuera plausible, todavía el eliminativista que quiera justificar su tesis tendría que mostrar que la incapacidad o incorrección explicativa hace necesaria la eliminación de las entidades a las que tales explicaciones refieren.

Ahora bien, la única manera en que este importante paso ontológico puede ser dado es bajo la presuposición de que el significado de los términos que aparecen en las descripciones y explicaciones de la Psicología Popular viene fijado por las leyes y estructuras teóricas que presiden el discurso acerca de deseos y creencias. Este punto de vista semántico —que es el adoptado por Paul Churchland— caracteriza la referencia de términos teóricos como *neutrón* a través del conjunto de leyes y estructuras teóricas de la física contemporánea. Si la teoría que engloba tales leyes es falsa, la consecuencia inmediata es que los términos teóricos que aparecen en sus leyes están vacíos de contenido. Las entidades a las que refieren no existen. Este ingrediente semántico es el que permite en última instancia a Paul Churchland dar el salto eliminativista definitivo a nivel ontológico[8].

Si podemos argüir que este ingrediente semántico es un ingrediente erróneo (cf. Sterelny, 1990, 147-148 y Greenwood, 1991), entonces, como dije, incluso si las explicaciones de la Psicología Popular resultan ser falsas o inadecuadas, esto no nos obliga a abandonar la ontología de estados que esta Psicología presupone.

El argumento se plantea como sigue. Es obvio que podemos referirnos a objetos incluso en aquellos casos en que las propiedades que les

7. Aunque con matices y orientación diferentes, el mismo tipo de ideas puede encontrarse en Bennett (1991) y Margolis (1991).

8. Esta postura radical a nivel ontológico se ha visto suavizada en los últimos escritos de Paul Churchland, en donde ya no parece negar existencia a los referentes de términos como «creencia» o «deseo», sino que los considera más bien como instrumentos de uso cotidiano que debemos usar sólo en ese ámbito. Ahora es el paradigma sentencial de representación el que sufre las mayores críticas. Cf. Churchland, P. M. (1991).

atribuimos no les corresponden en absoluto. La historia de la ciencia está llena de ejemplos de este tipo:

> La astronomía ptolemaica estaba completamente equivocada en lo relativo a la naturaleza de los planetas y las estrellas. Pero los defensores de la astronomía geocéntrica estaban equivocados *en lo referente a los planetas* cuando escribían que se mueven alrededor de la Tierra; cuando escribían eso no estaban escribiendo acerca de nada. Con el paso de la astronomía geocéntrica a la heliocéntrica cambiamos de opinión acerca de la naturaleza de los planetas en lugar de empezar a creer cosas sobre ellos por primera vez (Sterelny, 1990, 148).

Si esto es así, podemos estar completamente equivocados con respecto a la naturaleza de deseos y creencias sin que esto quiera decir que los estados a los que nos referimos con esos términos necesariamente hayan de ser eliminados de nuestro inventario ontológico.

Parte del debate en torno al eliminativismo se centra así en la cuestión filosófica acerca de las relaciones entre lo psicológico y lo físico, tanto a nivel epistemológico como ontológico. Pero la defensa o rechazo de una tesis reduccionista no lo es todo en la valoración del futuro de la Psicología Popular. Así, por ejemplo, los argumentos que han hecho de Stephen Stich —hasta ahora [9]— una bandera del eliminativismo están pincipalmente relacionados con el problema de la individuación del contenido semántico de los estados intencionales y tienen una prolongación importante en la cuestión empírica acerca del carácter simbólico o no que una teoría de los fenómenos cognitivos ha de poseer. El advenimiento del paradigma conexionista ha jugado un papel esencial en el desarrollo de este segundo aspecto del debate eliminativista. Y, aunque la pregunta central es empírica, los argumentos filosóficos han corrido paralelos a la *utilización* comparativa de las dos hipótesis en competición —clásica y conexionista—. En lo que resta me ocuparé de exponer y analizar estos argumentos.

IV. UNA RAZÓN PARROQUIANA: STEPHEN STICH

La imposibilidad de establecer una tesis reductiva con respecto a las leyes y entidades de la Psicología Popular no es el único argumento esgrimido en favor del eliminativismo. Según Stephen Stich, es el carácter relativista, multidimensional y, en definitiva, parroquiano de la taxonomía intencional imperante en la Psicología Popular el que pone de manifiesto su inadecuación como referencia teórica en el desarrollo de una genuina *Ciencia* Cognitiva.

9. «Hasta ahora» significa hasta la publicación de Stich y Warfield (1993); Stich, en preparación. Estas dos publicaciones son la prueba de un cambio radical en la posición de Stich respecto de estas cuestiones. Véase más abajo.

Stich proporciona algunos ejemplos y resultados provenientes de ciertos experimentos psicológicos para demostrar el carácter problemático de la individuación de creencias llevada a cabo en el seno de la Psicología Popular. Uno de los ejemplos más famosos es el de la Sra. T. y el asesinato de McKinley. La Sra. T. vivió en sus días de juventud el asesinato de McKinley y se vió profundamente impresionada por ello. Con el paso del tiempo, la Sra. T. empieza a perder ciertas facultades, entre ellas, la de la memoria. Ya senil, su memoria está tan deteriorada que olvida cosas como que las personas asesinadas están muertas. No obstante, la Sra. T. recuerda perfectamente que McKinley fue asesinado y puede proferir «McKinley fue asesinado» sin ningún tipo de problemas, aunque los tenga en recordar que las personas asesinadas no están vivas. ¿Cree *realmente* la Sra. T. que McKinley fue asesinado?

Para contestar a la pregunta general sobre si aguien cree realmente algo, nosotros establecemos, según Stich, una comparación implícita, relativa al conocimiento que ese alguien tiene sobre el tema. En el caso de la Sra. T., sin embargo, ese «conocimiento colindante» es tan peculiar que no ayuda a determinar si la atribución de esa creencia a la Sra. T. está o no justificada. La atribución de creencias a sujetos pertenecientes a culturas exóticas o, simplemente, a sujetos con creencias y deseos muy diferentes a los nuestros es problemática; en un sentido similar toda atribución de creencias se construye siempre sobre la base de una relativización o comparación implícita con las creencias sostenidas por los sujetos que efectúan tal atribución. Así, si nosotros —sujetos normales— atribuimos la creencia «McKinley fue asesinado» a la Sra. T., lo haremos sobre la base de lo que *nosotros* —para quienes los asesinados están muertos— *creemos* que la Sra. T. cree. Pero tal atribución está contaminada por nuestra perspectiva, ya que no se basa en ninguna idea precisa acerca de lo que debe ser —para la Sra. T.— el olvido de la relación entre muerte y asesinato más la creencia de que McKinley fue asesinado.

En general, de acuerdo con Stich, la individuación del contenido de una creencia —la identidad de tal contenido— en la Psicología Popular está mediatizada por su papel inferencial y su relativización a las creencias del sujeto que la atribuye. En ciertas condiciones, sin embargo, ese papel se resquebraja y la *identidad* se reduce, simplemente, a *similaridad* de contenido. La propia noción de creencia presente en la Psicología Popular incorpora esa vaguedad e imprecisión, pero una Ciencia Cognitiva no tiene cabida para nociones borrosas como ésta, y, por tanto, argüirá Stich, han de ser eliminadas.

El problema que Stich está planteando es un problema de taxonomización intencional o, para decirlo más precisamente, de la ausencia de una taxonomización intencional adecuada en el seno de la Psicología Popular. Para saber cuándo los pensamientos concretos sostenidos por diferentes individuos son pensamientos del mismo *tipo*, necesitamos ciertos

criterios taxonómicos de *identidad* del contenido. Pero, de acuerdo con Stich, el problema reside en que tal *identidad* de contenido es imposible de establecer. Factores como el papel inferencial interno, los conocimientos colaterales o las propiedades referenciales de los estados intencionales —que pueden variar con respecto al *mismo* estado intencional— intervienen de forma conjunta en la individuación de creencias, deseos, etc., haciendo imposible establecer un *tipo* o parte «naturalmente aislable» del sistema cognitivo que pueda hacerse corresponder con el contenido de tales estados intencionales.

La única noción a la que se puede aspirar en el seno de la Psicología Popular es la —mucho más diluida— de *similaridad* del contenido. Pero una teoría *científica* de la mente no puede incorporar o construirse sobre esta noción de *similaridad* de contenido de un estado intencional, porque tal teoría —arguye Stich— no puede satisfacer el absolutamente necesario *principio de modularidad*, *i.e.*, el principio de que los estados intencionales son estados funcionales discretos que pueden individuarse de forma definitiva con respecto a sus propiedades causales.

El argumento de Stich se eleva así sobre dos pilares igualmente importantes. Uno es la falta de correspondencia entre la organización causal del sistema cognitivo desde un punto de vista científico y desde el punto de vista de la Psicología Popular [10]. El otro, la imposibilidad de taxonomizar los estados intencionales de la Psicología Popular en términos correspondientes a clases naturales. La conclusión es que, muy probablemente, la Psicología Popular es falsa y que las entidades de las que habla no existen. La metáfora computacional conexionista ha servido además de soporte empírico —o, al menos, así lo ha defendido Stich— para estas conclusiones eliminativistas. Es dentro de este ámbito en donde se han desarrollado de forma más clara las ideas que hemos esbozado. Por esta razón merece un tratamiento aparte.

V. LA *CONEXIÓN ELIMINATIVISTA*

Existe un importante debate en el ámbito de la Inteligencia Artificial sobre si la clase de mecanismos a la que nosotros pertenecemos y a la que

10. Los ejemplos típicos que Stich proporciona para justificar esta afirmación residen en los resultados provenientes de una serie de experimentos sobre la —supuesta— interrelación de la conducta verbal y no verbal en individuos tratados de insomnio. Cf. Storms y Nisbett (1970), Nisbett y Wilson (1977), Nisbett y Ross (1980). En ellos se defiende la tesis del control dual, *i.e.*, la tesis de que —contrariamente a la idea predominante en la Psicología Popular— las conductas verbales y no verbales *no* están internamente conectadas sino que responden a mecanismos completamente diferentes. Si tal correlación interna no existe, entonces la noción de creencia, individuada con respecto a sus efectos verbales *y* no verbales, *no* es consistente. Siendo ésta una de las tesis fundamentales de la Psicología Popular, su falsedad conlleva la falsedad de tal doctrina (cf. Stich, 1983, 230-242).

tendría que remitirse el proyecto de modelado computacional de los procesos cognitivos está mejor representada por los modelos clásicos o por los denominados modelos conexionistas (McClelland, Rumelhart y col., 1986; Smolensky, 1987, 1988; Fodor y Pylyshyn, 1988; Pinker y Prince, 1988; Clark, 1989; Ramsey, Stich y Garon, 1991).

Frente a los modelos clásicos de procesamiento serial en los que la información está codificada en base a reglas de carácter lingüístico, los modelos conexionistas o de procesamiento distribuido de información en paralelo (PDP) contienen toda la información necesaria para explicar una cierta relación input-output sin referencia a ninguna categoría semántica. Las relaciones causales entre las unidades que componen el sistema describen suficientemente cómo la información es procesada por la red. Pero como estas unidades no tienen una interpretación semántica directa, todas las computaciones pueden ser explicadas sin referencia al *contenido* de la información procesada.

No debe resultar extraño que sean los modelos conexionistas los que aporten el soporte empírico apropiado para defender las tesis típicas eliminativistas. Después de todo, otra forma de entender el eliminativismo, una forma que ha sido desarrollada fundamentalmente por Stephen Stich bajo el rótulo de *Teoría Sintáctica de la Mente*, se basa en la idea de que los estados cognitivos pueden proyectarse sistemáticamente sobre objetos sintácticos abstractos, de tal manera que las cadenas causales entre estímulos y eventos conductuales pueden describirse exclusivamente en términos de propiedades y relaciones sintácticas de estos objetos sin necesidad de apelar a ningún contenido semántico.

La polémica entre los paradigmas clásico y conexionista en Inteligencia Artificial nos conduce así a un dilema: o bien claudicamos ante la posición eliminativista, si las hipótesis conexionistas son correctas —como aconsejan Ramsey, Stich y Garon (1991)—, o defendemos la eficacia causal de los estados mentales, *en tanto que estados con contenido semántico*, mostrando que, después de todo, las redes conexionistas nunca podrán ser realmente buenos modelos cognitivos (Davies, 1991).

Si optamos por el cuerno pro-conexionista, la idea intuitiva de que las actitudes proposicionales funcionan como causas de nuestra conducta parece ser una idea falsa. Las creencias o deseos, en tanto estados funcionalmente discretos, no son ni individualizables como estados de activación del sistema, ni individualizables en relación con sus efectos, porque la información es codificada por la red en representaciones distribuídas y superpuestas —véase más abajo—. Este es básicamente el argumento desarrollado por Ramsey, Stich y Garon.

Efectivamente, los mecanismos responsables de las relaciones *input-output* en los modelos conexionistas son las unidades que conforman la red. Cada unidad contribuye a codificar información acerca de diferentes proposiciones —esto es lo que significa que las representaciones son su-

perposicionales— y cada representación está así *distribuida* a lo largo de un conjunto amplio de micro-propiedades que tienen un carácter *subsimbólico*, *i.e.*, que no son semánticamente interpretables en sí mismas [11]. La producción de un determinado output supone la existencia de un determinado patrón de actividad de las unidades, aunque no necesariamente el mismo para representaciones del mismo tipo.

Puesto que las computaciones realizadas por la red están completamente determinadas al nivel de las unidades, y estas unidades codifican al mismo tiempo información de tipo diferente, no es posible identificar una entidad estable y recurrente que respalde la noción clásica de símbolo y que esté sujeta a las manipulaciones de un sistema de procesamiento independiente. Por otra parte, puesto que la misma información puede ser representada por redes con unidades y fuerzas de conexión diferentes, la clase de esas redes —posiblemente indefinida— no es más que un «conjunto caóticamente disyuntivo» que no corresponde en absoluto a la clase natural que resultaría, de acuerdo con la Psicología Popular, al considerar el conjunto de agentes cognitivos que tiene una creencia determinada.

La conclusión eliminativista que Ramsey, Stich y Garon establecen a propósito de estas consideraciones es que no tiene ningún sentido preguntar si la representación de una proposición determinada juega o no un papel causal en las computaciones de la red, porque no existe un estado discreto que corresponda a la información contenida en tal proposición, ni nada que sea comparable a una clase natural en el sentido psicológico. Si los fenómenos psicológicos no forman clases naturales, no pueden sustentar leyes causales universales que puedan ser en principio reducidas a leyes neurofisiológicas. Y, si esto es así, según Ramsey, Stich y Garon, nos vemos forzados a concluir que las entidades que tienen las propiedades esenciales tradicionalmente atribuidas a los fenómenos psicológicos por la Psicología Popular, no existen.

1. *Desurdiendo la trama. Niveles de descripción y unificación semántica*

El primer paso en la refutación de estos argumentos es una concesión al eliminativismo: si el nivel de análisis adoptado en la explicación de la conducta de los modelos conexionistas se reduce al nivel de las unidades del sistema —lo que Smolensky denomina nivel neurológico—, entonces las conclusiones eliminativistas son perfectamente plausibles. Pero, y éste es el punto importante —y el centro del argumento anti-eliminativista desarrollado por Andy Clark (cf. Clark, 1990)—, un tratamiento adecuado del paradigma conexionista, en tanto que *modelo cognitivo*, ha

11. Véase el capítulo sobre conexionismo en este mismo volúmen.

de desarrollarse a un nivel de descripción superior al de las meras activaciones numéricas de las unidades y las fuerzas de conexión.

Una de las primeras lecciones aprendidas en Ciencia Cognitiva es la de la existencia de múltiples niveles de descripción con respecto a un modelo computacional. La elección de uno u otro de estos niveles impone fuertes restricciones en el tipo de explicación que podemos proporcionar de la conducta del sistema. No es extraño, por tanto, que, limitándonos al nivel de fuerzas y unidades, obtengamos explicaciones poco satisfactorias desde el punto de vista de una conducta *semánticamente* interpretada.

Sin embargo, dentro de este paradigma existen técnicas de análisis que permiten ascender a un nivel de descripción superior, un nivel de descripción bajo el cual se pueden identificar representaciones y transformaciones que cumplen el mismo papel que las representaciones y reglas de los sistemas clásicos —si bien tienen un carácter completamente diferente a los símbolos y algoritmos clásicos, ya que tales representaciones y reglas son implícitas y altamente distribuidas.

Una de estas técnicas es la denominada análisis de grupo (*cluster analysis*). La idea básica de este método de tipo estadístico es la de extraer regularidades en los patrones de actividad que presentan las unidades ocultas para cada una de las relaciones *input-output* y utilizarlas para construir representaciones que agrupan relaciones con patrones similares. Esta técnica permite así unificar bajo categorías semánticas definidas patrones de actividad que, al nivel de las unidades y conexiones, tienen una estructura interna diferente. Y, lo más importante, permite establecer esta unificación semántica en función de *outputs* producidos a partir de *inputs* diferentes, *i.e.*, permite agrupar lo que al nivel de unidades son diferentes estados de una red en base a su eficacia causal respecto a una conducta determinada.

La misma técnica de análisis de grupo muestra que la parte del argumento eliminativista que ponía en cuestión la existencia de clases naturales, en el sentido en que esta noción se entiende en Psicología, es innecesariamente reduccionista. El hecho de que distintas redes puedan representar la misma información a pesar de estar constituidas por unidades con actividad y fuerzas de conexión diferentes deja de ser un problema si el análisis de tales redes se desarrolla a un nivel de descripción donde los *bloques* explicativos no son tanto las unidades cuanto sus patrones de actividad. Puesto que esos patrones resultan de la agrupación de mecanismos físicos diferentes que producen, sin embargo, los mismos outputs, el análisis del sistema a este nivel de descripción superior nos permite unificar lo que antes parecía un «conjunto caóticamente disyuntivo» en la misma clase de equivalencia (cf. Clark, 1990, *passim*)[12].

12. Ideas similares pueden encontrarse en Heil (1991) y Dennett (1991). En ambos artículos se pone de manifiesto cómo el carácter distribuido de las representacines descritas por los modelos cone-

2. *Stich* contra *Stich*

El propio Stich (Stich y Warfield, 1993) ha desarrollado recientemente un argumento en donde muestra que el condicional que deriva la verdad de la postura eliminativista a partir de la corrección de la hipótesis conexionista es falso, *i.e.*, el condicional cuya verdad él mismo había sostenido a través de los argumentos mencionados más arriba. Lo hace, además, criticando el contraargumento ofrecido por Clark [13].

Según Stich y Warfield, para que el condicional mencionado sea verdadero es necesario que sean verdaderos dos condicionales más:

a) si ciertos enfoques conexionistas son correctos, entonces la Psicología Popular está seriamente equivocada y

b) si la Psicología Popular está seriamente equivocada, entonces el eliminativismo es correcto.

Y, aunque se sigue manteniendo que (a) es verdadero —de ahí la crítica a Clark—, se cuestiona que (b) lo sea.

La crítica a la verdad del segundo condicional se realiza en dos movimientos, uno de los cuales ya ha sido analizado aquí. El primer movimiento consiste en poner de manifiesto cómo una eliminación a nivel ontológico de objetos tales como creencias y deseos sólo es posible si se defiende una teoría de la referencia para esos términos basada en una definición funcional implícita, de la manera en que vimos que la mantenía Paul Churchland. Ahora bien, como también vimos, la plausibilidad de este tipo de enfoque es altamente dudable. Stich y Warfield nos recuerdan sus fallos cuando se aplica a términos teóricos ya desaparecidos y muestran cómo el eliminativista de nuestro condicional (*b*), si quiere hacerlo verdadero, se encuentra en la muy difícil situación de tener que defender una teoría de la referencia basada en descripciones en contra del —mucho más adecuado— enfoque histórico-causal.

El segundo movimiento se basa en la crítica a la noción de propiedad constitutiva o necesaria de un objeto. Según Stich y Warfield, lo que el eliminativista necesita para hacer funcionar su argumento es mostrar que, para tener actitudes proposicionales, es necesario tener una propiedad que o bien la ciencia, o un argumento filosófico, puede probar que nadie tiene. Así, si para poder decir de alguien que tiene creencias y deseos debemos atribuirle constitutiva, *i.e.*, *necesariamente*, la propiedad «tener estados psicológicos modulados proposicionalmente», si podemos demos-

xionistas no entra necesariamente en conflicto con el papel causal *modular* tradicionalmente atribuido a las actitudes proposicionales.

13. De hecho, este artículo —Stich y Warfield, 1993— es una respuesta a los argumentos desarrollados tanto por A. Clark como por P. Smolensky en contra de la verdad de tal condicional. Cf. Clark (1990) y Smolensky (1987, 1988). No entraré sin embargo en los detalles de esta confrontación teórica. El dato relevante aquí es el giro que representa esta nueva línea con respecto a la tradicional postura eliminativista representada por Stich.

trar que nadie posee esa propiedad, tendremos razones para concluir que no existen cosas tales como creencias o deseos.

Ahora bien, no sólo la cuestión empírica sobre qué propiedades son necesarias para poder tener creencias y deseos es una cuestión abierta, también la cuestión de la existencia de propiedades conceptualmente necesarias es altamente problemática. Y ningún eliminativista ha aportado argumentos que faciliten o aclaren ninguna de las dos cuestiones.

Otra forma de expresar las mismas ideas es la siguiente. El Stich de (Ramsey, Stich y Garon, 1991) defiende que la Psicología Popular incorpora como propiedad necesaria de las actitudes proposicionales la modularidad proposicional, *i. e.*, la idea de que las actitudes proposicionales «son estados funcionalmente discretos y semánticamente interpretables que juegan un papel causal en la producción de otras actitudes proposicionales y, en definitiva, en la producción de conducta» (Ramsey, Stich y Garon, 1991, 204). Esta propiedad, sin embargo, no es atribuible a las entidades que intervienen en las redes conexionistas, porque tales entidades tienen un carácter distribuido y subsimbólico. Luego, si estos modelos conexionistas son modelos correctos de la cognición humana, entonces nos enfrentamos a un cambio teórico *ontológicamente radical* con respecto a esas peculiares entidades teóricas que son las actitudes proposicionales, *i.e.*, hemos de reconocer que no existe *nada* que tenga esa propiedad y, por tanto, abandonar cualquier tipo de compromiso con respecto a la *existencia* de las actitudes proposicionales.

El nuevo giro representado por Stich y Warfield defiende, sin embargo, una visión de las relaciones entre conexionismo y Psicología Popular enmarcada en un cambio teórico *ontológicamente conservador*. Se acepta que, si los modelos conexionistas son correctos, hemos de admitir que la modularidad proposicional no se encuentra entre las propiedades que podemos atribuir a las actitudes proposicionales, pero, se argumenta, ello no nos lleva directamente a prescindir en nuestra ontología de tales entidades. De esta manera el «segundo Stich» parece desligarse completamente de su posición eliminativista anterior y acercarse al tipo de contraargumentos expuestos al final de la sección III.3. Este *arrepentimiento* teórico no se produce exclusivamente en la versión computacional de su doctrina, sino que alcanza además sus fundamentos más estrictamente filosóficos: los problemas de individuación de contenido semántico y desajuste causal. *Deconstructing the Mind* (Stich, en preparación) promete ser una elaboración provocativa y, sin duda, interesante de este nuevo periodo.

VI. PARA RESUMIR

El instrumento artesanal que constituye la Psicología Popular fue caracterizado en la introducción como un cuerpo explicativo y predictivo de la conducta cotidiana. Lo que define típicamente a este tipo de explicaciones y predicciones es el recurso a estados mentales que pueden formularse en términos proposicionales a la hora de establecer las causas de la conducta —el recurso a cosas tales como deseos, creencias, sospechas, temores, etc.

La tesis eliminativista, por el contrario, establece que la teoría que resulta de la estructuración de los principios básicos de la Psicología Popular es una teoría estancada, pobre, inadecuada y, en definitiva, falsa y que, por tanto, las entidades que componen el universo teórico de la Psicología Popular han de ser desestimadas como inexistentes.

Si la tesis eliminativista fuera verdadera, el edificio completo de la Psicología Popular parecería derrumbarse, arrastrando consigo todo un repertorio ontológico de actitudes básicas para nuestra vida diaria. El análisis tanto de los argumentos que la sustentan como de los correspondientes contraargumentos ha sido nuestro principal objetivo. Los argumentos han sido divididos en cuatro grupos diferentes.

El primero de ellos, recogido en la sección II, y desarrollado por Paul Churchland, mantiene como tesis básica la incapacidad y estancamiento explicativos de la Psicología Popular para con fenómenos psicológicos centrales —como el sueño, la imaginación o las enfermedades mentales—. A esta tesis responden Horgan y Woodward con dos líneas argumentativas diferentes. La primera defiende que los fenómenos señalados por Paul Churchland quedan fuera del rango de fenómenos que el aparato teórico de la Psicología Popular puede siquiera generar. La segunda pone de manifiesto que las estrechas relaciones entre Psicología Cognitiva y Psicología Popular hacen más difícil la acusación de estancamiento para esta última e insiste en la inadecuación de la idea de progreso como criterio evaluativo de la Psicología Popular.

La sección III recoge los argumentos encuadrados dentro de la línea reduccionista y diferencia los de Paul y Patricia Churchland (secciones III.1 y 2). Mientras Paul aboga por la eliminación de nuestro inventario ontológico de aquellas entidades que forman el entramado teórico de la Psicología Popular, la línea de Patricia Churchland, más comedida —aunque igualmente reduccionista—, es una crítica dirigida más específicamente contra el modelo sentencial de representación mental, el modelo que tiene su pilar básico en la hipótesis fodoriana del Lenguaje del Pensamiento.

El denominador común de las respuestas a este ataque reduccionista consiste en la maniobra teórica de aceptar la irreductibilidad de la Psicología Popular a neurociencia, pero negar que tal reducción sea condi-

ción necesaria de su verdad. Dentro de esta maniobra se encuadran los contraargumentos de Horgan y Graham en términos de la noción realista de sistema intencional resonante, el enfoque instrumentalista de Dennett, así como la crítica al componente semántico subyacente a la posición de Paul Churchland. Todos ellos contribuyen a mostrar la debilidad de la postura reduccionista y aparecen recogidos en la sección III.3.

La tercera categoría de argumentos en nuestra clasificación tiene como protagonista principal a Stephen Stich y su peculiar teoría sobre individuación y atribución de creencias basada en la *similaridad* de contenido semántico. De acuerdo con Stich (sección IV), es esta noción débil de similaridad de contenido la única que la Psicología Popular puede ofrecer en el marco de su taxonomización intencional. Ahora bien, el problema consiste en que una teoría que pretenda ser *científica*, no puede construirse sobre una noción tan débil, porque ello supondría la imposibilidad de satisfacer el denominado *principio de modularidad*, *i.e.*, el principio que defiende el carácter discreto de los estados intencionales en virtud de sus propiedades causales.

La versión computacional de esta misma idea aparece desarrollada en la sección V y constituye la cuarta y última categoría taxonómica. Esta línea, defendida por Ramsey, Stich y Garon, está basada en la defensa de la verdad del siguiente condicional: si el conexionismo es correcto, entonces la tesis básica de la Psicología Popular de que son las actitudes proposicionales las que funcionan como causas de nuestra conducta es una idea falsa. En otras palabras, si aceptamos que los modelos conexionistas son buenos modelos cognitivos, hemos de aceptar igualmente que la tesis eliminativista es verdadera. La razón reside, según Ramsey, Stich y Garon, en que el carácter distribuido y superpuesto de las representaciones en los sistemas conexionistas hace imposible la determinación de aquellos estados discretos que actúan como causas y que constituyen las clases naturales de la Psicología Popular.

Las secciones V.1 y 2 están dedicadas a desarrollar dos intentos diferentes de refutar esta línea argumentativa. En la primera, se defiende la compatibilidad conceptual y taxonómica de la Psicología Popular y el modelo conexionista recurriendo a un método estadístico de partición del espacio representacional. Esta partición proporciona una taxonomía paralela a la de los constituyentes de las actitudes proposicionales de la Psicología Popular.

La segunda —que tiene como peculiaridad haber sido desarrollada por el propio Stich— sigue manteniendo la existencia de una relación necesaria entre la corrección del conexionismo y la incorrección de la Psicología Popular, pero pone de manifiesto que la eliminación de las entidades correspondientes al espacio ontológico de la Psicología Popular necesita de algo más que de esa incorrección. Ese algo más pasa por una revisión del componente semántico presente en las tesis eliminativistas así

como por un replanteamiento de la noción de propiedad necesaria de un objeto.

En general, el embate eliminativista no parece sobrevivir a los argumentos desarrollados en su contra. Esto es especialmente cierto del aspecto ontológico de la tesis, en donde la eliminación de entidades como deseos y creencias resulta difícil de sostener. Cuando el objetivo de la crítica lo representa más bien el modelo sentencial de representación mental, el paso ineludible consiste en mostrar la conexión necesaria entre la Psicología Popular y tal modelo de representación. Ahora bien, demostrar la existencia de tal conexión interna no es asunto sobre el que argumento filosófico o descubrimiento empírico alguno haya sentado una tesis definitiva. El tema no está cerrado, pero mientras Psicología Cognitiva y Neurociencia siguen trabajando y se siguen elaborando nuevas posiciones filosóficas, el futuro de la Psicología Popular no parece tan amenazado como sus enemigos eliminativistas nos habían hecho creer.

BIBLIOGRAFÍA

Baker, L. R. (1987), *Saving Belief*, Princeton University Press, Princeton.

Bennett, J. (1991), «Folk Psychological Explanations», en J. D. Greenwood (ed.), 1991, 176-195.

Blackburn, S. (1991), «Losing Your Mind: Physics, Identitiy and Folk Burglar Prevention», en J. D. Greenwood (ed.), 1991, 196-225.

Clark, A. (1989), *Microcognition. Philosophy, Cognitive Science and Parallel Distributed Processing*, MIT Press, Cambridge, Mass.

Clark, A. (1990), «Connectionist Minds»: *Proceedings of the Aristotelian Society, vol. XC*, 83-102.

Clark, A. (1993), «The Varieties of Eliminativism»: *Mind and Language*, en prensa.

Cortázar, J. (1970), «Tía en Dificultades», en Id., *Historias de Cronopios y Famas*, Edhasa, Barcelona, 39-40.

Churchland, P. M. (1979), *Scientific Realism and the Plasticity of Mind*, Cambridge University Press, Cambridge.

Churchland, P. M. (1981), «Eliminative Materialism and the Propositional Attitudes»: *The Journal of Philosophy, 78*, 67-90. Reimpreso en W. G. Lycan (ed.), 1990, 206-223.

Churchland, P. M. (1986), «Some Reductive Strategies in Cognitive Neurobiology»: *Mind, 95*, 279-309.

Churchland, P. M. (1988), *Matter and Consciousness*, MIT Press, Cambridge, Mass.

Churchland, P. M. (1990), *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*, MIT Press, Cambridge, Mass.

Churchland, P. M. (1991), «Folk Psychology and the Explanation of Human Behavior», en J. D. Greenwood (ed.), 1991, 51-69.

Churchland, P. S. (1980), «A Perspective on Mind-Brain Research»: *Journal of Philosophy, 77*, 185-207.

Churchland, P. S. (1981), «Language, Thought and Information Processing»: *Nous, 14*, 147-170.
Churchland, P. S. (1986), *Neurophilosophy*, MIT Press, Cambridge, Mass.
Davies, M. (1991), «Concepts, Connectionism and the Language of Thought», en W. Ramsey, S. Stich y D. Rumelhart (eds.), 1991, 229-257.
Dennett, D. (1987), *The Intentional Stance*, MIT Press, Cambridge, Mass.
Dennett, D. (1991), «Two Contrasts: Folk Craft *versus* Folk Science, and Belief *versus* Opinion», en J. D. Greenwood (ed.), 1991, 135-148.
Eckardt, B., von (1984), «Cognitive Psychology and Principled Skepticism»: *The Journal of Philosophy, 81*, (2), 67-88.
Feyerabend, P. (1963), «Materialism and the Mind-Body Problem»: *Review of Metaphysics, 17*, 49-67.
Fodor, J. (1975), *The Language of Thought*, Thomas Crowell, New York.
Fodor, J. (1988), *Psychosemantics*, Bradford/MIT Press, Cambridge, Mass.
Fodor, J. y Pylyshyn, Z. (1988), «Connectionism and Cognitive Architecture»: *Cognition, 28*, 3-71.
Goldman, A. (1989), «Interpretation Psychologized»: *Mind and Language, 4*, 161-185.
Gordon, R. (1986), «Folk Psychology as Simulation»: *Mind and Language, 1*, 158-171.
Graham, G. y Horgan, T. (1988), «How to be Realistic about Folk Psychology»: *Philosophical Psychology, 1*, 69-81.
Greenwood, J. D. (ed.) (1991), *The Future of Folk Psychology. Intentionality and Cognitive Science*, Cambridge University Press, Cambridge.
Greenwood, J. D. (1991), «Reasons to Believe», en Id. (ed.), 1991, 70-92.
Heil, J. (1991), «Being Indiscrete», en J. D. Greenwood (ed.), 1991, 120-134.
Horgan, T. y Graham, G. (1991), «In Defense of Southern Fundamentalism»: *Philosophical Studies, 62*, 107-134.
Horgan, T. y Woodward, J. (1985), «Folk Psychology is Here to Stay»: *The Philosophical Review, 94*, (2), 197-226. Reimpreso en W. G. Lycan (ed.), 1990, 399-420, por donde se cita.
Jackson, F. y Petit, P. (1990), «In Defence of Folk Psychology»: *Philosophical Studies, 59*, 31-54.
Kitcher, P. (1984), «In Defense of Intentional Psychology»: *The Journal of Philosophy, 81*, (2), 89-106.
Lycan, W. G. (ed.) (1990), *Mind and Cognition. A Reader*, Basil Blackwell, Oxford.
Margolis, J. (1991), «The Autonomy of Folk Psychology», en J. D. Greenwood (ed.), 1991, 242-262.
MacDonald, G. y MacDonald, C. (1993), *Folk Psychology*, Blackwell, Oxford, en prensa.
McCauley, R. (ed.) (en preparación), *The Churchlands and Their Critics*, Blackwell, Cambridge.
McClelland, J., Rumelhart, D. y col. (1986), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1 y 2, MIT Press, Cambridge, Mass.
McGinn, C. (1989), *Mental Content*, Basil Blackwell, Oxford.

Nisbett, R. y Ross, L. (1980), *Human Inference Strategies and Shortcomings of Social Judgement*, Prentice-Hall, Englewood Cliffs, NJ.
Nisbett, R. y Wilson, T. (1977), «Telling More than We Know: Verbal Reports on Mental Processes»: *Psychological Review, 84*, 231-279.
Pinker, S. y Prince, A. (1988), «On Language and Connectionism. Analysis of a Parallel Distributed Processing»: *Cognition, 28*, 73-193.
Quine, W.V. (1966), «On Mental Entities», en Id., *The Ways of Paradox and Other Essays*, Random House, New York.
Ramsey, W., Stich, S. y Garon, J. (1991), «Connectionism, Eliminativism and the Future of Folk Psychology», en W. Ramsey, Stich y D. Rumelhart (eds.), 1991, 199-228.
Ramsey, W., Stich, S. y Rumelhart, D. (eds.) (1991), *Philosophy and Connectionist Theory*, Lawrence Erlbaun, London.
Rorty, R. (1965), «Mind-Body Identity, Privacy and Categories»: *Review of Metaphysics, 19*, 24-54.
Smolensky, P. (1987), «Connectionism, AI and the Brain»: *Artificial Intelligence Review, 1*, 95-109.
Smolensky, P. (1988), «On the Proper Treatment of Connectionism»: *Behavioural and Brain Sciences, 11*, 1-74.
Sterelny, K. (1990), *The Representational Theory of Mind. An Introduction*, Basil Blackwell, Oxford.
Stich, S. (1982), «On the Ascription of Content», en A. Woodfield (ed.), 1982, 153-205.
Stich, S. (1983), *From Folk Psychology to Cognitive Science*, MIT Press, Cambridge, Mass.
Stich, S. y Warfield, T. (1993), «Do Connectionist Minds Have Beliefs? A Reply to Clark and Smolensky», en G. MacDonald y C. MacDonald (eds.), 1993, en prensa.
Stich, S. (en preparación), *Deconstructing the Mind*. Manuscrito.
Storms, M. y Nisbett, R. (1970), «Insomnia and the Attribution Process»: *Journal of Personality and Social Psychology, 16*, (2), 319-328.
Woodfield, A. (ed.) (1982), *Thought and Object*, Clarendon Press, Oxford.

EVOLUCIÓN Y LENGUAJE

Antoni Gomila Benejam

I. INTRODUCCIÓN

Al pretender abordar, una vez más, la cuestión del origen del lenguaje, resulta inevitable empezar justificando tal pretensión frente a la prohibición de tratar el tema, por acientífico, decretada por la Sociedad Lingüística de París, en sus estatutos fundacionales, en 1866. En realidad, nada más fácil. Por una parte, porque el interés del tema sigue siendo enorme. Por otra, porque los avances en nuestro conocimiento ocurridos en el siglo largo transcurrido desde entonces, permiten abordar la cuestión desde presupuestos muy distintos. Vayamos por partes.

Para empezar, basta con tener presente lo central que resulta el lenguaje para nuestra propia autocomprensión como humanos: desde Descartes hasta Turing, por señalar dos posiciones extremas con respecto a la naturaleza de la inteligencia humana, la capacidad para el lenguaje ha sido y es considerada como marca distintiva de lo mental, como índice inequívoco de humanidad. Comprender el origen del lenguaje supone, por ello, avanzar en la comprensión de nuestra naturaleza humana, de nuestra mente, de lo que constituye la base de nuestra especificidad. Más allá, cualquier reconstrucción plausible del origen del lenguaje ofrecerá necesariamente respuestas a cuestiones relacionadas con la organización de nuestras capacidades mentales: la relación lenguaje-pensamiento, la relevancia funcional de nuestra dimensión social, la naturaleza del desarrollo humano, la organización funcional del cerebro... Inversamente, el ámbito del origen del lenguaje puede servir como «base de pruebas» de las diversas teorías centradas en cada uno de estos temas, al requerir su consistencia y apoyo interteórico. Como ha señalado Bickerton:

> La evolución del lenguaje es tan reciente que podemos suponer razonablemente que su naturaleza actual está todavía condicionada por esos orígenes, y su papel crucial en la distinción entre nosotros y las demás especies (...) es tal que debe influir fuertemente, aunque no determine completamente, todo lo que pensamos y hacemos (Bickerton, 1981, 315).

Desde esta perspectiva, resulta clara la necesidad de plantearse de nuevo la cuestión del origen del lenguaje. Por su propia naturaleza —se trata de reconstruir un proceso que tuvo lugar hace ya tiempo, y que no dejó registro directo—, nuestra cuestión requiere un tratamiento interdisciplinar, para recoger las múltiples evidencias, aunque indirectas, ofrecidas desde las diversas disciplinas que tienen algo que aportar al estudio de la mente y el lenguaje. Y es evidente que los avances y variedad de propuestas teóricas disponibles en los diversos campos que pueden tener relevancia para nuestra cuestión son muy distintos de lo que eran hace un siglo. La Lingüística y la Psicolingüística, la Pragmática y la Filosofía del Lenguaje y de la Mente, la Etología y la Sociobiología, la Paleontología y la Antropología, las Neurociencias... aportan ideas y conocimientos indudablemente superiores a los de hace un siglo. Pero lo más importante es el cambio fundamental que se ha producido en la consideración del hombre: la consolidación de la biología evolucionista en este siglo ha situado todas las investigaciones sobre nuestra especie en un plano esencialmente distinto del que predominaba en 1866. Las consecuencias de este cambio para la cuestión del origen del lenguaje son decisivas: nuestra cuestión se ha transformado en la cuestión de la filogénesis del lenguaje, de su dimensión evolutiva, y el enfoque para tratarla, por tanto, no tiene por qué diferir del adoptado con respecto a cualquier otra característica evolutiva.

Sin embargo, sería incorrecto decir que los múltiples avances experimentados apuntan todos en la misma dirección, o que de ellos emerge una respuesta homogénea a nuestra cuestión. Al contrario, por una parte hay disputas teóricas dentro de cada una de las disciplinas; por otra, es posible más de una reconstrucción de la filogénesis del lenguaje en base al conocimiento disponible, y por supuesto, hay aspectos todavía oscuros. Curiosamente, no obstante, en el momento presente el mayor problema, a nuestro entender, no radica en el «cuándo» del proceso, sino en el «porqué», no tanto en las etapas del proceso, cuanto en el aspecto conceptual, en formular la lógica del proceso de la filogénesis, el mecanismo responsable de la periodización del proceso. Ya desde las primeras propuestas al respecto, las de Condillac, éste es el principal problema al que hacer frente. Para Condillac (1746), el lenguaje es una invención humana; pero ¿cómo puede inventarse algo así y transmitirse a los demás sin disponer ya de un lenguaje? Una circularidad semejante se presenta en propuestas contemporáneas (Bennett, 1976; Searle, 1979; Tennant,

1984), que pretenden explicar la aparición de significados lingüísticos convencionales en términos de la teoría de Lewis (1969). Esta teoría presupone una intencionalidad de segundo orden, la capacidad de pensar sobre los propios pensamientos e intenciones, en los sujetos participantes, como base para la fijación convencional de significados. Pero ¿cómo es posible alcanzar ese alto grado de capacidad simbólica e intencional independientemente del lenguaje?

Algo similar ocurre cuando se consideran los aspectos involucrados en el desarrollo del lenguaje, en particular, la necesidad de postular un dispositivo innato específico, defendido convincentemente por la Psicolingüística de inspiración chomskiana. Dado que la estimulación necesaria para activar este mecanismo debe ser ya lingüística (no basta cualquier sonido), uno se pregunta cómo pudo aparecer en primer lugar tal mecanismo específico, si cuando no había todavía lenguaje (por lo que no sería específico del lenguaje), o si cuando ya había lenguaje (por lo que no sería necesario para el lenguaje) (Bickhard, 1979).

Pero este tipo de problemas —problemas de qué fue antes, el huevo o la gallina—, tienen de hecho solución en el marco evolutivo: un huevo y una gallina más primitivos (o no específicos). En este trabajo vamos a intentar bosquejar los mecanismos del proceso de aparición del lenguaje de forma no circular, tanto en los aspectos filogenéticos como ontogenéticos, en base a los prerrequisitos necesarios para la aparición de cada nueva capacidad en su complejidad, sin prestar tanta atención a los problemas de periodización de las fases del proceso. Para ello comenzaremos por establecer la naturaleza de aquello cuya filogénesis nos interesa: qué características se precisan para que un medio comunicativo pueda ser considerado del mismo tipo que nuestro lenguaje. Esto nos permitirá valorar apropiadamente los resultados provenientes de la experimentación con animales, en especial con primates, así como obtener una primera idea de la complejidad de los mecanismos mentales que posibilitan el lenguaje. A continuación consideraremos los aspectos relacionados con el desarrollo lingüístico en el niño, ya que constriñen de forma muy precisa el proceso filogenético, por las implicaciones innatistas que conllevan. Seguidamente, nos centraremos en los requisitos necesarios para el lenguaje a nivel de mecanismos y estructuras neurofisiológicas y anatómicas (tipo de características del cerebro, del aparato fonador).

Cada una de estas tres dimensiones —qué es el lenguaje, cómo se desarrolla y qué estructuras precisa— contribuye a delimitar nuestro problema y sus posibles explicaciones; y conjuntamente, caracterizan lo que podríamos llamar el estadio final del proceso filogenético. El siguiente apartado consiste en el núcleo de nuestra propuesta: una reconstrucción del escenario evolutivo que permita aclarar el conjunto de cambios involucrados en la aparición del lenguaje. La clave del proceso, a nuestro modo de ver, consistió en la nueva forma de organización so-

cial derivada del cambio de nicho ecológico, y la coevolución de capacidad representacional-capacidad lingüística, vía aumento de la capacidad cerebral, como forma de respuesta a las nuevas presiones selectivas. La conclusión final intentará recoger el tema de las consecuencias que pudo tener el impacto de la aparición del lenguaje en relación con el pensamiento reflexivo y la autoconciencia.

II. DE LA COMUNICACIÓN AL LENGUAJE

En tanto que sistema vocal de comunicación, el lenguaje consiste en un medio de emparejar significados (o mensajes) con sonidos, y viceversa (ya que cubre tanto los aspectos de producción como de comprensión). Sin embargo, presenta características especiales en cada uno de esos niveles: los significados son conceptualmente complejos, los sonidos se organizan según patrones de articulación, y la forma de emparejarlos resulta de la combinación sistemática de unidades correspondientes a una serie jerárquica de niveles. Además, el uso de este sistema es independiente de los estímulos presentes, por lo que se incorporan también aspectos pragmáticos en su funcionamiento apropiado [1].

Comenzando por el nivel semántico, puede decirse que la base de la funcionalidad del lenguaje como medio de comunicación reside justamente en su gran potencial simbólico, en su capacidad para expresar gran diversidad de significados. Esta riqueza de nuestros recursos conceptuales se refleja en el nivel léxico, pero sobre todo en la estructura proposicional, sujeto-predicado, de los significados. De este modo, los mensajes lingüísticos consisten en atribuir propiedades a individuos o clases de individuos, o relaciones entre ellos. El significado del mensaje, según una semántica composicional, es función de los significados de las unidades léxicas que participan más las reglas que regulan sus combinaciones. Gracias a esta estructura y a la riqueza léxica, los mensajes no se refieren necesariamente al momento y lugar en que se produce la emisión —ni siquiera cuando se trata de deícticos su contenido se agota en la situación de emisión—, y pueden incluir la negación entre sus recursos simbólicos.

Esta potencial productividad —no hay un límite al número de mensajes expresables— y sistematicidad —el significado de una expresión depende de los significados de las unidades léxicas constituyentes y de las reglas de su combinación— del nivel semántico establece fuertes limitaciones a la forma en que puede darse el emparejamiento con los sonidos que expresan esos contenidos. La capacidad humana de producir y dis-

1. Vamos a limitarnos a una caracterización general de los niveles del lenguaje. Un tratamiento más específico puede hallarse en Akmajian, Demers y Harnish (1984).

tinguir una gran variedad de sonidos está bien documentada, pero es fácil ver por qué un emparejamiento uno a uno, entre significados y sonidos, no sería una buena solución: al escoger el emisor un nuevo sonido para expresar un nuevo significado, se encontraría la dificultad de que el receptor no reconociera esa nueva relación. De ahí la necesidad de contar con un modo sistemático de generar las expresiones sonoras a partir de las representaciones conceptuales. Esto es lo que hace el nivel sintáctico: permite representar los significados en forma de cadena sonora, mediando entre las propiedades estructurales de los significados y las secuenciales de los sonidos, al compartir propiedades con ambos tipos de representaciones, en base a una serie de categorías (complemento, núcleo...), de relaciones (dominancia, precedencia) y de principios (suyacencia, proyección) (Newmeyer, 1991).

Esta combinación de unidades léxicas según reglas estructurales constituye la base de la generatividad del lenguaje, o infinitud discreta, según expresión de Chomsky, que permite expresar la riqueza del nivel semántico. Con una propiedad añadida, que garantiza que el nivel sintáctico refleje la sistematicidad y productividad del nivel semántico, la recursividad, esto es, que las reglas sintácticas puedan ser aplicadas sucesivamente a su propio output, sin otro límite que la propia capacidad de procesamiento del sujeto. En definitiva, el nivel sintáctico asegura la posibilidad de recuperar un nuevo mensaje a partir de una nueva secuencia sonora, ya que la novedad no es arbitraria, sino sistemática.

El propio nivel fonológico (y su realización fonética) se encuentra también organizado sistemáticamente, en base a unidades elementales (los fonemas) y sus posibilidades de combinación, dependientes de las propiedades del aparato fonador humano y la capacidad de control motor para articular los sonidos. De hecho, es preciso un mecanismo al menos tan rico como nuestro sistema sonoro para expresar nuestros complejos contenidos proposicionales. Nuestro sistema de escritura, por ejemplo, los tiene, pero justamente porque imita, si bien no en detalle, los medios expresivos sonoros. Es una cuestión abierta, a la que volveremos, si otros medios representacionales pueden permitir la misma riqueza expresiva.

Finalmente, es preciso incluir un nivel específicamente pragmático en el lenguaje. Como ya hemos mencionado, el uso lingüístico no depende de la presencia de ciertos estímulos, sino que es esencialmente voluntario. Esta propiedad de la actividad lingüística permite separar el significado de la actitud emocional del emisor —se puede hablar del miedo sin estar temblando—, hacer uso de los mensajes con intenciones diversas (declarativa o interrogativa, exclamativa o rogativa, etc.), así como realizar acciones esencialmente lingüísticas (prometer, insultar, maldecir, confesar...) En todos estos casos, la actividad lingüística se sitúa en un nivel intencional reflexivo, en el que se toman en cuenta no sólo las propias in-

tenciones comunicativas, sino también la situación mental de la audiencia a la que uno se dirige (sus estados de conocimiento o ignorancia, de atención o interés,...) y el contexto.

Davidson (1982) ha sostenido, en base a estos aspectos pragmáticos, que sólo puede hablarse propiamente de lenguaje cuando se dispone de esta capacidad para la conciencia reflexiva, para pensar sobre los propios pensamientos, y por ello, de voluntad comunicativa, e inversamente, que sólo quien tiene lenguaje puede de hecho pensar. Se sigue como consecuencia que sólo los hombres tenemos propiamente lenguaje. En realidad, esta afirmación de la exclusividad humana del lenguaje se sigue no sólo de las características específicas del uso del lenguaje, sino que no es atrevido afirmar, en nuestra opinión, el carácter únicamente humano del lenguaje en base a cualquiera de las propiedades de los otros niveles, semántico, sintáctico o fonológico. Para mostrarlo, podemos considerar uno de los sistemas de comunicación vocálica más sofisticado que encontramos en la naturaleza, el de los monos vervets (Seyfarth y Cheney, 1990).

Estos monos tienen tres tipos distintos de vocalización, con significados diferentes: uno se emite ante la detección de águilas; otro, ante la detección de serpientes, y otro, por fin, ante la detección de leopardos. Parece tratarse, pues, de un lenguaje, dado que se trata de tres pares significado-sonido. Las diferencias con nuestro lenguaje, sin embargo, son grandes. A nivel conceptual, los recursos conceptuales de estos monos son muy limitados: por ejemplo, no pueden referirse al águila de la semana pasada, ni referirse a los tres tipos de depredadores como casos del concepto depredador, ni atribuir propiedades o relaciones a los objetos que detectan, ni mucho menos negar que posean propiedades. Su repertorio, además, es muy limitado, por lo que la asociación sonido-significado es directa, sin necesidad de una estructura sintáctica que permita la generatividad de expresiones. A nivel fonético, los gritos de los monos vervets no son articulados, por lo que, aunque esta cuestión no ha sido estudiada en particular, cabe suponer que su repertorio sonoro debe de ser también muy limitado. Por último, las limitaciones pragmáticas son evidentes: las llamadas son elicitadas por la presencia de estímulos característicos; a su vez, la llamada da lugar a ciertas reacciones apropiadas típicas por parte de la audiencia (de subir a los árboles si la llamada corresponde a leopardo, de refugiarse bajo los árboles si corresponde a águila, de ponerse en pie y mirar entre las hierbas si corresponde a serpiente). No puede distinguirse entre mensajes imperativos y declarativos, ni tampoco separar el mensaje del estado emocional del emisor. Y aunque sólo se emiten las llamadas si se está próximo al propio grupo, se ha comprobado que no corrigen a las crías cuando emiten incorrectamente alguna señal, lo que indica un uso no intencional de su código.

Todos estos aspectos sitúan este sistema comunicativo en un plano cualitativamente distinto del lenguaje humano: mientras en el caso hu-

mano es posible distinguir entre el contenido del mensaje y los efectos que el mensaje produce en el receptor, ya que el mensaje sólo puede tener consecuencias si es entendido, y esas consecuencias pueden variar según el receptor, en el caso de los monos, la comunicación puede comprenderse desde el análisis causal, esto es, como un medio de influir en la conducta ajena (Krebs y Dawkins, 1984). Los signos de estos sistemas de comunicación son el resultado de un proceso de ritualización (Lorenz, 1977; Wilson, 1975) por el que ciertos sonidos (o conductas, o rasgos morfológicos) son exagerados, o repetidos, o esquematizados, de forma tal que se convierten en indicadores fiables de ciertas condiciones externas (la presencia de un depredador, por ejemplo), por lo que se convierten en sustitutos del propio estímulo en la producción de la conducta apropiada a ese estímulo. De ahí que se afirme que, en el fondo, estos sistemas comunicativos no son propiamente simbólicos, ya que sus signos no son arbitrarios, convencionales, sino que son icónicos, aunque la relación original del signo con lo significado se haya perdido en el proceso de ritualización (Brandon y Hornstein, 1986).

Puede alegarse, sin duda, que los monos vervets no son el mejor candidato que podría presentarse para ilustrar las mejores capacidades animales. Ciertamente se han conseguido resultados más espectaculares con primates superiores, pero significativamente, siempre en situaciones no naturales [2]. Así, está claro que las capacidades conceptuales de los chimpancés son elevadas. Como han mostrado sobre todo los estudios de Premack (1985, 1988) y Savage-Rumbaugh (Savage-Rumbaugh, 1986; Savage-Rumbaugh et al., 1986), los chimpancés son capaces de usar signos arbitrarios, y pueden referirse con ellos a objetos distantes en el tiempo o el espacio. Se ha documentado incluso el caso de un chimpancé enano (Pan Paniscus) capaz de entender ciertas expresiones del inglés.

Sin embargo, la evidencia disponible parece indicar un grado muy limitado de generatividad. Recuérdese la expresión de Washoe —«agua pájaro»— al ver por primera vez un cisne (Gardner y Gardner, 1969), y otras parecidas, pero que pueden ser interpretadas igualmente como una simple asociación de ideas (Terrace, 1979). Más seguros son los resultados obtenidos por Herman con delfines (Herman, Richards y Wolz, 1984), los cuales, tras aprender a obedecer instrucciones expresadas mediante secuencias de hasta cinco signos, consiguieron responder adecuadamente a secuencias nuevas. Esta comprensión de una nueva secuencia indicaría la capacidad de representarse el modo de generar secuencias significativas. No obstante, esta capacidad de los delfines es limitada y claramente no supone recursión. Curiosamente, el caso más claro de generatividad no

2. Snowdon (1982) constituye una revisión de los estudios de la comunicación vocal en los primates, concluyendo que las a menudo largas secuencias de llamadas son altamente estereotipadas, semejando más el canto de los pájaros que el lenguaje, aunque pueden discernirse ciertas regularidades.

proviene de un primate, sino de un pájaro (*Parus Atricapillus*). Hailman y Ficken (1986) han identificado cuatro elementos sonoros básicos, que pueden combinarse indefinidamente según ciertas reglas, en sus cantos. Por ejemplo, las series ABBCD, AAACDDD, o BBBCCDD son «correctas», mientras que las series BAACD, AAACCDC o BBBDDDCC no lo son. Sin embargo, las reglas de estos cantos pueden ser expresadas mediante una gramática de estados finitos, sin recursión, ya que las opciones válidas en cada momento dependen del elemento precedente. Y, en cualquier caso, tales series no parecen expresar contenidos proposicionales.

En cuanto al nivel fonético, ya descubrió Yerkes en los años veinte la incapacidad articulatoria de los chimpancés. De ahí que se les hayan enseñado otros tipos de signos: signos del lenguaje de los sordomudos, tarjetas, lexigramas... que permitan generar diversidad de expresiones. La ausencia de nivel sintáctico, sin embargo, lleva a que el aprendizaje se produzca por emparejamiento de cada signo con cada significado, lo que conlleva a su vez la incapacidad de producir nuevas expresiones sígnicas. Finalmente, parece que puede hablarse de comunicación proto-intencional en los chimpancés, y por tanto, de un cierto nivel pragmático, si bien las limitaciones de su código dificultan que se pueda diferenciar entre mensajes imperativos y declarativos, por ejemplo (Gómez, Sarriá y Tamarit, 1993).

En resumen, como han afirmado Seidenberg y Pettito, «la dicotomía entre las capacidades cognitivas y las capacidades lingüísticas de los monos es el hallazgo más importante que ha emergido de la moderna investigación con monos» (1987, 284). Lo que tiene una implicación directa para nuestra investigación: la aparición del lenguaje debe haberse producido después de que los homínidos se separaran de la línea evolutiva de los monos. Y además, que la clave en la aparición del lenguaje es el nivel sintáctico, aunque probablemente tuviera consecuencias para el propio nivel simbólico. Lo que deberemos hacer es reconstruir cuáles fueron los factores que impulsaron la evolución homínida, partiendo de una situación en que ya se disponga de capacidades cognitivas del mismo tipo, al menos, que las de los chimpancés[3].

3. Por lo que sé, sólo Hurford (1987) discrepa de la tesis de que el lenguaje sea un fenómeno exclusivamente homínido, en base a la posibilidad de que la capacidad lingüística permanezca latente en los primates. Esta opinión, no obstante, presupone que lo que han aprendido los chimpancés de los experimentos es lenguaje. Más frecuente es la opinión de que, siendo el lenguaje un producto propiamente homínido, existe continuidad y no ruptura con el resto de los primates, trazando el origen del lenguaje en ciertas capacidades, comunicativas o no, ya presentes en los primates, como programas motores (Lieberman, 1984), gestos comunicativos (Bates, 1976; Greenfield y Smith, 1976), o en mecanismos dedicados a la conceptualización de la topología (Talmy, 1988). En el otro extremo, los seguidores de Chomsky afirman la radical discontinuidad evolutiva que supone el lenguaje (Piatelli-Palmarini, 1989; García-Albea, 1993). Pero ésta puede convertirse fácilmente en la típica discusión bizantina: en todas las novedades evolutivas hay continuidad y ruptura. De nuevo, resulta fundamental distinguir entre convergencia y homología evolutiva.

III. INNATISMO LINGÜÍSTICO Y DESARROLLO DEL LENGUAJE

Una dificultad importante, que hemos señalado ya en la introducción, para intentar esa reconstrucción filogenética radica en la naturaleza del proceso ontogenético de desarrollo del lenguaje. En efecto, como ha puesto claramente de manifiesto la Psicolingüística de inspiración chomskiana, y más concluyentemente, la Teoría Formal del Aprendizaje, el niño no aprende el lenguaje —en el sentido de generalizar las reglas sintácticas a partir de los estímulos que le llegan—. La complejidad gramatical es tal que sin correcciones, sin datos negativos, no es posible alcanzar las reglas de una gramática de los lenguajes humanos, y esas correcciones no se dan (Grimshaw y Pinker, 1989; Marcus, 1993), a pesar de lo cual el lenguaje se aprende. Además, los estímulos lingüísticos con que entra en contacto el niño son construcciones elementales, que no ejemplifican la potencial complejidad gramatical (Morgan, 1989). Por otra parte, se adquiere el lenguaje independientemente de la capacidad de aprendizaje o coeficiente intelectual (síndrome de Down, por ejemplo), y de tener o no déficits sensoriales (ciegos, sordos...) (Mehler y Dupoux, 1990). Es preciso, por tanto, concluir que el niño aporta una parte importante al proceso de adquisición, que se ha dado en llamar dispositivo para la adquisición del lenguaje (DAL), un mecanismo específico para el lenguaje.

Por otra parte, hace falta una estimulación específicamente lingüística para que el DAL se active, no sirve cualquier estímulo sonoro. De lo cual se deduce que el cerebro del niño se encuentra ya predispuesto al lenguaje, predispuesto a captar las regularidades que rigen los usos lingüísticos del grupo en que se encuentra. Consistente con esta conclusión es el hecho de que exista un período crítico en el desarrollo del lenguaje, un límite temporal a su adquisición (alrededor de los diez años) (Newport, 1990). Estas dos características ponen de relieve, de hecho, la importancia del proceso de desarrollo infantil en la conformación de nuestras capacidades mentales.

Al tomar en consideración esta dimensión de desarrollo, nuestro problema se vuelve mucho más preciso: no se trata solamente de explicar cómo pudo aparecer el lenguaje por primera vez, sino de explicar también cómo pudo aparecer el componente innato que resulta necesario para la adquisición del lenguaje actualmente. Y ambas cuestiones parecen ir en direcciones opuestas: cuando apareció el lenguaje, no podía haber ya un mecanismo específico para el lenguaje; y si pudo aparecer sin ese mecanismo, ¿por qué se desarrolló un dispositivo de este tipo? Evitar esta paradoja requerirá distinguir entre un primer momento de aparición de un protolenguaje con una cierta estructura, sin unos mecanismos especializados en su producción y comprensión, pero cuya funcionalidad favoreciera la aparición de tales mecanismos específicos, según principios evolutivos generales, lo que a su vez llevaría a una mayor eficiencia en la

adquisición de la competencia lingüística infantil. Lo que nos interesa, en este punto, es plantearnos lo que es innato del lenguaje, lo que el niño «nace sabiendo» (Mehler y Dupoux, 1990).

De especial relevancia para esta cuestión resultan dos casos de anomalías de desarrollo: el lenguaje de los niños sordomudos de nacimiento y las lenguas criollas. En ambos casos encontramos un patrón parecido: una sintaxis más pobre de lo normal en la primera generación (parecida a la de las *lingu franc* desarrolladas para el comercio con Oriente durante la Edad Media), pero un lenguaje completo en la segunda; en definitiva, encontramos ejemplos de creación de lenguaje. Un grupo de niños sordomudos de padres hablantes desarrollan espontáneamente un código de signos [4] con una estructura elemental; pero en la siguiente generación puede hablarse ya de un lenguaje con el mismo grado de complejidad sintáctica que las lenguas habladas [5] (Klima y Bellugi, 1979).

Lo mismo encontramos en el caso de las lenguas criollas, desarrolladas en el siglo pasado por los esclavos de las plantaciones coloniales. Una historia poco conocida, quizá, pero que constituyó uno de los mayores experimentos sociales de la historia. El origen étnico de los esclavos de una plantación (en Hawai o Jamaica, en Madagascar o la Guayana) no era homogéneo, al contrario, y podían encontrarse diversidad de lenguas en una misma región. En situaciones de este tipo aparecen entonces primero una lengua de contacto llamada «pidgin», con predominio léxico del idioma de los esclavistas (inglés o francés, según las zonas) pero con una estructura propia, distinta y pobre. En cambio, en la siguiente generación, los niños desarrollan ya una lengua, la criolla, sintácticamente completa (Bickerton, 1984).

La creatividad lingüística, además, no está limitada a casos anómalos como los citados, sino que está presente en todos los niños. Como ha señalado Bickerton (1990), estos casos encuentran difícil acomodo en la concepción psicolingüística predominante del DAL como Gramática Universal, en particular, como un conjunto de principios generales y una serie de parámetros que deben ser fijados, en función del contexto lingüístico en que se encuentre el niño (Lightfoot, 1989). Estos casos indican claramente la capacidad de ir más allá de la estructura dada en la estimulación lingüística con que entra en contacto el niño, y sugiere, por

4. Nótese que hablamos de signos —o señas— y no de gestos para poner de manifiesto el carácter simbólico del código. Esto es de especial relevancia para quienes defienden, con Condillac, el origen gestual del lenguaje humano (Hewes, 1973): también los chimpancés gesticulan, pero su capacidad para aprender un lenguaje de señas, como el Ameslan, como hemos visto en relación con los trabajos de Premack, es muy limitada. Más grave resulta, incluso, considerar estos lenguajes como evidencia del origen gestual del lenguaje (Corballis, 1992)

5. Sacks (1989) ha descrito brillantemente los avatares históricos de los lenguajes de sordomudos, la incomprensión y represión con que frecuentemente han topado, y casos históricos de algunas comunidades predominantemente sordomudas, incluyendo una universidad.

tanto, una concepción del DAL como un mecanismo facilitador, en lugar de como la caracterización abstracta de todos los lenguajes posibles.

Estos datos del desarrollo lingüístico, además, encajan bien con la idea de que el DAL surgió como una especialización funcional para facilitar el acceso a un medio de importancia vital. Cabe esperar de ello, por consiguiente, predisposición a desarrollar cierto tipo de estructuras primeramente, en lugar de una especificación de las propiedades abstractas de todo lenguaje. En realidad, quienes conciben el DAL como Gramática Universal se ven abocados a reconocer que tal caracterización universal alcanza sólo a lo que llaman el «núcleo» de las lenguas, reconociendo la diversidad de las «periferias» (Chomsky, 1981), que deberían ser aprendidas. La relación entre la GU y el DAL, sin embargo, no tiene por qué ser tan directa: las propiedades formales de las lenguas pueden ser una consecuencia de las reglas facilitadoras del proceso de desarrollo lingüístico, sin que ello presuponga que tales propiedades deban estar en el niño[6].

En este punto, no obstante, es preciso plantearse una última cuestión: ¿por qué no es innato todo el lenguaje? Es concebible, sin duda, una situación en que el lenguaje fuera algo parecido a la visión, un mecanismo general cuyo funcionamiento apropiado requiere solamente un periodo de afinación, pero sin necesidad de estimulación específica o de aprendizaje. Habría, claro está, en tales circunstancias, un solo lenguaje, que el niño empezaría a hablar muy pronto.

Resolver satisfactoriamente esta cuestión no es sencillo, pero su planteamiento nos remite de nuevo a la naturaleza única del desarrollo infantil humano, y nos lleva a plantearnos el sentido evolutivo de este desarrollo, que abordaremos de forma directa en la sección V. La idea básica consiste en reconocer los límites de la preprogramación genética como vía adaptativa cuando las condiciones en que hay que desenvolverse son cambiantes; los procesos de desarrollo, por el contrario, pertenecen al nivel epigenético (cuando el fenotipo no está totalmente determinado por el genotipo y puede ser sensible, por tanto, a condiciones ambientales diversas), por lo cual permiten mayor flexibilidad e indican variabilidad ambiental (Plotkin y Odling-Smee, 1979). Que el lenguaje tenga que desarrollarse, entonces, responde de nuevo a la importancia de la creatividad lingüística, ya inicialmente, y a la esperable diversidad de lenguas que se sigue de ella. Lo cual no impide, no obstante, que la ne-

6. Hemos defendido esta concepción alternativa del DAL, justamente en base a la naturaleza evolutiva del lenguaje, en Gomila (en prensa), y en relación a una concepción más abstracta de la relación entre competencia gramatical y representación mental en Gomila (1992). Dado que el desarrollo de las ideas en torno a un componente innato en la competencia lingüística ha tenido lugar al margen de consideraciones evolutivas, resultaría ciertamente sorprendente que éstas apoyaran sin más las concepciones psicolingüísticas de inspiración chomskiana desarrolladas.

cesidad de un periodo de desarrollo relativamente largo (o, inversamente, un prolongado periodo de inmadurez) fuera resultado de procesos previos a la aparición del lenguaje (cf. sección siguiente).

En resumen, la consideración de la ontogenia del lenguaje nos ha servido para precisar mejor las condiciones que nuestra hipótesis acerca de su filogenia debe satisfacer y ha ofrecido algunas claves a ese respecto (por ejemplo, que la articulación sonora no es el único medio de expresar la riqueza conceptual de nuestros mensajes). Necesitamos ahora incluir en nuestro cuadro las estructuras neuroanatómicas que sostienen la capacidad lingüística y lo que revelan en relación a la historia del lenguaje.

IV. REQUISITOS ESTRUCTURALES PARA EL LENGUAJE

En cierto sentido, la consideración de las condiciones estructurales para el lenguaje ofrece una base más segura para determinar nuestra cuestión que los aspectos funcionales del lenguaje. A primera vista, si conseguimos establecer qué mecanismos anatómicos son imprescindibles para el lenguaje, podremos establecer de forma fiable la aparición del lenguaje en base a la aparición de tales mecanismos en la filogénesis humana. Este es, de hecho, el razonamiento que está a la base de los primeros intentos de recuperar la cuestión del origen del lenguaje en épocas recientes (Harnad y cols., 1976). Así, se puede considerar que la clave para el lenguaje es la aparición de un área de asociación intermodal en el cerebro (Smith, 1985; Wilkins y Dumford, 1990), y por tanto, fijar así la aparición del lenguaje. O bien puede considerarse que lo esencial es la lateralización funcional del cerebro (Gazzaniga, 1983), o la presencia de áreas específicas para el lenguaje (Broca, Wernicke) y tomar entonces la aparición de tales características cerebrales como criterio (Tobias, 1987). Igualmente, a nivel del aparato fonador, puede establecerse que la capacidad articulatoria del sonido exige determinada posición de la laringe, y decidir el origen del lenguaje en base a reconstrucciones de las estructuras del aparato respiratorio de los homínidos (Lieberman, 1984).

Lamentablemente, estas estrategias distan de resultar satisfactorias, y mucho menos concluyentes. En primer lugar, por la dificultad de contar con reconstrucciones fiables de los cerebros y demás estructuras no óseas de los homínidos, a pesar de los grandes avances ocurridos. Así, por ejemplo, Lieberman (1984) concluye que los Neanderthales eran incapaces de articular los sonidos, por la posición de su laringe según su reconstrución. Pero, como ha señalado Falk (1975), si la reconstrucción fuera correcta, entonces también serían incapaces de tragar, lo que resulta inaceptable. Además, este enfoque pasa por alto la posibilidad, ya mencionada, de un lenguaje de señas. En segundo lugar, esas estructuras criteriales resultan no serlo tanto, ya que parecen encontrarse también en

los simios tanto la lateralización hemisférica (Bradshaw y Nettleton, 1989), como el área de asociación intermodal (Pinker y Bloom, 1990), sin que se dé capacidad lingüística, como hemos visto. Pero el problema de fondo con estos planteamientos radica en el razonamiento de partida: identificar la estructura que actualmente permite realizar cierta función no garantiza que esa misma estructura realizara esa misma función en el pasado. Es posible tanto que ya existiera la función, mediante otra estructura, como que esa estructura desempeñara otra función[7]. Y esas posibilidades son especialmente significativas cuando se trata del cerebro. De todos modos, aun en el caso, ciertamente deseable, de que fuera posible identificar la aparición de la estructura cerebral responsable del lenguaje, ello no nos suministraría ipso facto una explicación evolutiva del lenguaje: nuestra cuestión consiste en comprender por qué surgió una tal estructura.

La conclusión que se sigue de todo ello no es, por supuesto, que tales investigaciones carecen de interés; más bien, que su relevancia debe ser juzgada a la luz de otras evidencias, conductuales y funcionales. De mayor interés resultan, por ello, los estudios que ponen de relieve lo que podríamos llamar dependencias estructurales en los primates, las proporciones entre diversos aspectos anatómicos, o anatómicos y conductuales, ya que permiten detectar desviaciones ostensibles y significativas en el caso humano, indicando tanto la presencia de presiones selectivas como de constricciones estructurales para las diversas vías adaptativas.

La primera y más obvia medida de este tipo es el índice de encefalización. Como ya notó Darwin en *The Descent of Man*, no es el tamaño del cerebro lo significativo en el hombre, sino su gran tamaño en relación al tamaño del cuerpo. Como ha mostrado Jerison (1976), nuestro cerebro es tres veces más grande de lo que cabría esperar según la proporción normal entre los primates —que ocupan ya el extremo superior del proceso de encefalización que caracteriza a los mamíferos—. Lo distintivo de la filogénesis humana radica justamente en este aumento progresivo del índice de encefalización en el género Homo. Este aumento cerebral, además, no es homogéneo, sino que ciertas áreas experimentaron un desarrollo mucho mayor que otras, destacando especialmente el córtex prefrontal, donde se localizan justamente las funciones lingüísticas (Deacon, 1988).

Curiosamente, este aumento del cerebro va acompañado de un cambio en la cara, que pasa de estar proyectada hacia adelante a ocupar una posición ortogonal con respecto al cerebro, amén de otros cambios: reducción de los arcos supraciliares y de los caninos, la recesión de la

7. Sin embargo, resulta significativo el descubrimiento de una zona cerebral especializada en el reconocimiento de llamadas vocales de miembros de la propia especie en muchos mamíferos, que corresponde aproximadamente a nuestra área de Wernicke (Bradshaw y Nettleton, 1989). La cuestión, en cualquier caso, sigue siendo si se trata de un caso de homología o de convergencia.

mandíbula y la formación de la barbilla. La razón de estos cambios colaterales parece relacionada con otro índice de proporción anatómica en los primates, el de la relación entre la cantidad de materia ósea en relación al peso total del cuerpo, un 6 o 7 % (Potter, 1986). Al aumentar la necesidad de materia ósea craneal (para cubrir un cerebro mayor), habría disminuido la disponibilidad de materia ósea para conformar otras estructuras, que habrían experimentado una reducción. De hecho, se ha observado esta correlación entre aumento del cerebro y disminución de la cara, y viceversa, en los macacos (Albrecht, 1978). Igualmente, la poderosa mandíbula inferior del Australopitecus Boisei va acompaña de un frágil y fino cráneo (Rak, 1987).

El interés de esta dependencia estructural radica en que la recesión del rostro derivada del aumento craneal constituiría el factor desencadenante del descenso de la laringe, clave para la formación de nuestro específico sistema articulatorio. Al retroceder el rostro, disminuye la cavidad oral y con ello, el espacio para la lengua, vital para el proceso mecánico de engullir (Biggerstaff, 1977). Con el descenso de la laringe se habría conseguido mayor espacio para la lengua, al poder doblarse hacia atrás, en lugar de estar plana en la boca.

Estos cambios, inicialmente motivados por el aumento cerebral, pudieron tener a su vez consecuencias con respecto al tipo de alimentación apropiada: la gracilidad mandibular impide una alimentación basada en raíces y hojas —abundante, pero que exige gran ingestión para satisfacer la necesidades proteínicas—, en favor de una alimentación centrada en semillas y frutas (dependiente por tanto de factores estacionales), y ocasionalmente carne —alimentos más ricos en proteínas pero que hay que buscar en un territorio más extenso, con los cambios consiguientes en la organización social (Clutton-Brock y Harvey, 1980; Foley, 1987; Milton, 1993).

En este sentido, es preciso tener en cuenta otro índice característico de los primates: el que correlaciona tamaño del cerebro y tamaño del grupo (Dunbar, 1993), lo que indica un aumento sostenido en el tamaño de los grupos homínidos hasta alcanzar alrededor de ciento cuarenta miembros en el Homo Sapiens. En tales circunstancias, la importancia selectiva de un medio para establecer y preservar la estabilidad del grupo, asegurando la cooperación y permitiendo la enculturación de los nuevos miembros, no puede ser pasada por alto.

Este aspecto nos remite a un último indicador significativo de nuestra evolución, el relacionado con la inmadurez de los recién nacidos humanos. Si siguiéramos el patrón primate, la gestación humana debería durar 18 meses (Krogman, 1972). Sin embargo, ello haría imposible el parto, por las limitaciones físicas del canal pélvico, derivadas a su vez de nuestra condición bípeda. La «solución», pues, consiste en el nacimiento prematuro, que nos sitúa de nuevo ante la importancia del proceso de de-

sarrollo en la conformación del bebé, constituyendo la base de la flexibilidad cognitiva humana. Su viabilidad depende, sin duda, de garantizar la protección y cuidados precisos durante el largo período de dependencia absoluta de la madre, así como de pautas apropiadas de integración social. Es en tales circunstancias que resulta fundamental disponer de un medio de transmisión de información simbólica (conocimientos y valores sociales) (Brandon y Hornstein, 1986). Ésta es el marco selectivo que puede dar cuenta de la evolución del lenguaje.

V. ENCEFALIZACIÓN Y LENGUAJE: UN ESCENARIO EVOLUTIVO

Aunque nos hemos mantenido en un nivel de generalidad, contamos ya con los mimbres principales para intentar elaborar la estructura, al menos, de nuestro cesto. Como señalamos en la introducción, nuestra reconstrucción de la filogénesis del lenguaje ha de enmarcarse en el marco evolutivo general. Quizá merezca la pena detenernos un instante para considerar la forma de este tipo de explicaciones.

En un proceso evolutivo cabe distinguir cuatro tipos de factores: condiciones ambientales, causas (presión selectiva), constricciones (límites mecánicos, energéticos, ontogenéticos... a la profundidad del cambio evolutivo) y consecuencias (Foley, 1990). En ciertas condiciones, la presencia de una presión selectiva favorece un cierto tipo de cambio adaptativo (aparición de nuevas estructuras, modificación de otras preexistentes...), que puede verse limitado por constricciones (ciertas organizaciones biológicas son físicamente imposibles) y dependencias estructurales (hay que partir de lo disponible); sin embargo, si la presión es suficientemente fuerte, las constricciones pueden verse superadas permitiendo, de este modo, una modificación más radical, y llevando a una reorganización de otros aspectos. Las consecuencias, tanto a nivel de reorganización como de repercusiones conductuales, pueden alterar a su vez cada uno de los factores involucrados, lo que puede llevar a nuevos cambios. El proceso evolutivo, así considerado, es un proceso dinámico, donde no hay «adaptaciones» en sentido absoluto, sino siempre relativo a ciertas condiciones y constricciones [8]. Esta dimensión del proceso filogenético en que las consecuencias de un cambio funcional afectan a su

8. Desde esta perspectiva, pierde fuerza la distinción introducida por Gould y Vrba (1982) entre adaptaciones —estructuras resultado de la selección natural— y «exaptaciones» (o, en terminología darwiniana, pre-adaptaciones) —estructuras inicialmente sin función que pasan a tenerla, o que cambian de función—. Al ver el proceso evolutivo en perspectiva dinámica, la distinción se vuelve borrosa, o mejor, la exaptación deja de aparecer como opuesta a la adaptación para resultar un tipo de adaptación: la selección natural nunca parte de cero, opera siempre en ciertas condiciones y con ciertas constricciones, no como un diseñador o ingeniero; de ahí su oportunismo, la continuidad estructural que permite la aparición de nuevas funciones (Jacob, 1977).

vez a ese cambio recibe el nombre de coevolución (Dawkins, 1986). Como veremos, resulta fundamental a nivel teórico para poder dar cuenta de la diversidad de cambios involucrados en la aparición del lenguaje.

Como ilustración, tomemos el caso de las alas de los pájaros. Una explicación satisfactoria, aunque imprecisa, de su origen como estructura para volar consiste en señalar una función anterior, en este caso termoreguladora, para los antecedentes evolutivos de la estructura actual. En algún momento, un cambio en las condiciones ecológicas convirtió en más útiles unas estructuras termorreguladoras que otras: las que permitían, además de calentar el cuerpo, desplazarse en el aire, y así, ocupar nuevos nichos ecológicos. La selección natural, entonces, llevó a una modificación gradual de esas estructuras termorreguladoras de tal forma que pudieran llevar a cabo mejor su nueva función. Este proceso, además, no es cerrado en sí mismo: puede suponer a su vez cambios en el regimen alimenticio del pájaro o en su patrones conductuales que, a su vez, pueden reforzar el desarrollo de formas de vida aéreas.

Contra este enfoque funcional o adaptativo de la evolución del lenguaje, Chomsky (1980, 1982, 1988), entre otros[9], ha negado la posibilidad de que la selección natural pueda explicar el origen evolutivo del lenguaje. La complejidad y sofisticación del lenguaje, en su opinión, exceden con mucho lo que podría considerarse como utilidad biológica. Se trataría más bien de un maravilloso subproducto del proceso de selección de cerebros cada vez mayores, no sometido a su vez, sin embargo, a las condiciones de la selección natural. Sensible a las críticas de Gould y Lewontin (1979) al uso indiscriminado de las explicaciones funcionales, Chomsky sugiere que la capacidad lingüística resultaría de propiedades formales del cerebro o de efectos pleiotrópicos[10] a nivel genético:

> Estas habilidades [lingüísticas] bien pueden haber aparecido como concomitantes de propiedades estructurales del cerebro que se desarrollaron por otras razones. Supóngase que hubo selección de mayores cerebros, mayor superficie cortical, especialización hemisférica para el procesamiento analítico, o muchas otras propiedades estructurales que podemos imaginar. El cerebro que evolucio bien pudo tener todo tipo de propiedades especiales que no son seleccionadas individualmente; no habría nada milagroso en esto, sino solamente el funcionamiento normal de la evolución. No tenemos idea, en la actualidad, de cómo se aplican las leyes de la física cuando 10^{10} neuronas se sitúan dentro de un objeto del tamaño de una cesta, en las condiciones especiales que surgieron durante la evolución humana (Chomsky, 1980, 321).

9. Cf. Mehler (1985), Premack (1985), Piatelli-Palmarini (1989), García-Albea (1993).

10. En general, un gen no interviene solamente en un aspecto del fenotipo, sino en varios al mismo tiempo. Puede ser seleccionado por uno de estos aspectos, pero el resultado se manifiesta también en la retención de los demás aspectos. Este carácter múltiple de los efectos de los genes se conoce como pleiotropismo.

Sin embargo, y dejando aparte la ausencia de propuesta alguna acerca de cómo podría derivarse el lenguaje de otras propiedades estructurales del cerebro, es preciso tener en cuenta que es justamente la complejidad de nuestra capacidad lingüística el mejor indicio de que estamos ante un producto de la selección natural, tal como ya señaló Darwin y han insistido recientemente Pinker y Bloom (1990). El mecanismo de selección natural es el único que nos permite explicar el fenómeno de la complejidad adaptativa, la «ilusión de diseño inteligente», la aparición de estructuras o conductas cuya naturaleza responde al desempeño de cierta función compleja. Casi cualquier cosa puede servir de pisapapeles, pero no como medio de comunicación. De hecho, el propio Chomsky presupone que el proceso de encefalización que caracteriza la filogénesis homínida responde a presiones selectivas; ¿por qué, entonces, rechazar que este proceso involucró también al lenguaje? Por el contrario, nuestra hipótesis consiste en que la evolución del lenguaje sólo tiene sentido en el contexto del proceso de encefalización [11].

En efecto, el proceso de encefalización constituye el aspecto más llamativo del proceso filogenético humano, y existen ideas bastante claras respecto a las condiciones ecológicas que lo favorecieron, así como de las constricciones que llevaron a que adoptara la forma en que lo hizo y de las consecuencias morfológicas, dietéticas y sociales a que dio lugar. Estas consecuencias, a su vez, modificaron las condiciones iniciales en tal forma que, por una parte, permitieron la aparición de un medio simbólico de comunicación, y por otra, convirtieron en altamente útil un medio de este tipo, generando a su vez un nuevo proceso evolutivo por el que se especializaron mecanismos cognitivos en el procesamiento y producción de lenguaje [12].

En primer lugar, pues, como las diversas medidas de proporciones neuroanatómicas revisadas en la sección anterior sugieren, la causa fundamental de la filogénesis homínida fue la presión selectiva en favor de un mayor cerebro en relación al córtex, hasta el punto de que la desvia-

11. El anti-funcionalismo extremo de Chomsky con respecto al lenguaje es utilizado por sus críticos para desacreditar su concepción del desarrollo ontogenético del lenguaje, en razón de que no encaja en el marco de la teoría de la evolución. Cf. Bates, Thal y Marchman (1989) y Lieberman (1984). Sin embargo, el necesario funcionalismo a nivel filogenético no implica la validez del funcionalismo a nivel ontogenético, como defienden estos autores, sino que más bien apoya el anti-funcionalismo innatista chomskiano. Téngase en cuenta, por ejemplo, que la competencia gramatical comienza a aparecer alrededor de los dos años, bastante antes de la competencia intencional (o teoría de la mente), que no aparece hasta los cuatro (Wimmer y Perner, 1983), a partir de capacidades intencionales más básicas (Bruner, 1983). Ello indica que el desarrollo ontogenético del lenguaje no responde a exigencias de la comunicación intencional.

12. Téngase en cuenta que nuestro escenario evolutivo se refiere al paso de seres con capacidades comunicativas y simbólicas parecidas a las de los primates superiores a seres con nuestras capacidades lingüísticas. Los cambios en las condiciones ambientales no afectan a todas las especies por igual: de ahí la necesidad del concepto de nicho ecológico cuya caracterización tiene lugar por referencia a la especie que lo ocupa.

ción con respecto a la proporción estructural esperable según el patrón primate es de tres veces. Encontramos una desviación paralela en el periodo de gestación, dada la constricción física del canal pélvico, consecuencia a su vez del bipedalismo, que conduce a un nacimiento prematuro, y la necesidad de un decisivo periodo de desarrollo.

Para entender el sentido adaptativo de este aumento cerebral, centrado particularmente en el córtex, es preciso tener en cuenta las condiciones ecológicas de nuestros antepasados homínidos: la ocupación del hábitat de la sabana tras la recesión del bosque tropical como consecuencia de la glaciación. Este nuevo hábitat presenta novedades importantes: mayor dispersión de las fuentes de alimentación, tanto espacialmente (necesidad de ocupar un territorio mayor) como en el tiempo (estacionalidad, ausente en el clima tropical), y al mismo tiempo mayor riesgo en el desplazamiento, por ser mayor y por ser, con frecuencia, a campo abierto. En tales circunstancias, resulta adaptativa la que podría llamarse estrategia de la flexibilidad fenotípica (Brandon y Hornstein, 1986): ser capaz de conformar la propia conducta a las condiciones relevantes. Cuando estas condiciones son impredecibles en términos del tiempo requerido para que puedan ser incorporadas al bagaje genético, no queda otra vía más que la del aprendizaje individual, la de la adquisición de información y conocimiento acerca de ese entorno variable y hasta cierto punto imprevisible: capacidad de orientarse al seguir rutas variables, de recordar dónde y cuándo pueden encontrarse ciertos frutos o semillas, de saber cuáles son comestibles y cuáles venenosos, de anticipar qué va a ocurrir en base a indicios fiables... Del mismo modo, es preciso que la propia conducta sea función de este conocimiento (conducta intencional), y pueda planificarse, en relación a propósitos compartidos, así como a vínculos y alianzas, relaciones de reciprocidad, etc.

Es fácil ver que el proceso de encefalización contribuye precisamente a desarrollar esta función adaptativa. El aumento del sistema nervioso central constituye, por una parte, la base para esa adquisición de conocimiento del entorno, y por otra, la base del control voluntario de la conducta, frente a reflejos e instintos. Precisamente, la inmadurez cerebral en el nacimiento constituye la clave neurofisiológica de esta flexibilidad adaptativa, al conformarse la estructura cerebral frente a frente de la experiencia individual, con lo que resulta sensible al entorno en el que este desarrollo individual tiene lugar.

Resulta instructivo, en este punto, detenernos para notar el carácter coevolutivo del proceso, con varios factores influyéndose mutuamente, apuntando en la misma dirección: un cambio en las condiciones ecológicas favorece un aumento cerebral como ajuste adaptativo a la novedad de un ambiente impredictible (a nivel genético), que tiene como consecuencia, en razón de constricciones mecánicas, un nacimiento prematuro, que favorece a su vez el proceso de adaptación a estas nuevas condiciones

al conllevar que la conformación cerebral sea sensible a las condiciones presentes en el proceso de desarrollo.

Sin embargo, hay que avanzar un paso más y considerar otras consecuencias de este proceso, que alteran a su vez las propias condiciones en que tiene lugar, favoreciendo con ello nuevos desarrollos que culminan en la aparición del lenguaje. A nivel social, resulta claro que la inmadurez del recién nacido exige asegurarle atenciones y cuidados máximos, en un primer momento, y además, asegurarle un proceso de aprendizaje eficaz y fiable del medio. Resulta fácil mostrar que en este sentido es más ventajoso un medio que permita transmitir la información acumulada por los progenitores que el medio de aprender esa información de nuevo por ensayo y error. Igualmente, la utilidad de contar con un medio de transmisión de información se incrementa en relación a la necesidad de organización del propio grupo, cuyo tamaño aumenta en relación al incremento del tamaño del cerebro, como hemos mencionado, lo que supone un cambio importante en su composición, ya no caracterizada por la consanguineidad de sus miembros, por lo que son precisas nuevas formas de mantener la cohesión y la cooperación en su seno.

Por otra parte, los cambios en la estructura ósea del cráneo requeridos para posibilitar el proceso de encefalización suponen, en primer lugar, una cierta fragilidad mandibular, que refuerza a su vez la dependencia de una dieta omnívora rica en proteínas, en lugar de una dieta basada en la ingestión de gran cantidad de fibra (raíces y hojas), menos dependiente de la estacionalidad y que requiere un menor territorio, con lo que consolida la estrategia de adquisición de información; además, como también se mencionó en la sección anterior, estos cambios óseos podrían ser la causa del descenso de la laringe, modificación destinada a preservar la capacidad de engullir, pero que ocasiona como consecuencia la posibilidad de articular sonidos.

En estas nuevas condiciones resulta clara la funcionalidad de un medio de transmisión de información simbólica. La estrategia evolutiva que caracteriza la filogénesis homínida consiste en la flexibilidad adaptativa, en la sensibilidad a un contexto abierto, resultado de la experiencia individual y no de la completa preprogramación genética (como veremos, es ventajoso anticipar desde el principio los aspectos robustos, omnipresentes, del ambiente), y la consiguiente mayor capacidad de actuar intencional y creativamente, todo ello fuertemente dependiente, al mismo tiempo, de condicionantes sociales: de una enculturación eficaz y fiable, del mantenimiento de la cohesión y cooperación intragrupal, aspectos que favorecen el desarrollo de un medio de comunicación simbólica. Un medio, por otra parte, que no resulta ya fuera del alcance de este estadio evolutivo. Aunque con limitaciones, se dispone ya de algunas de las condiciones necesarias para la comunicación oral intencional: voca-

lización articulada (probablemente acompañada de señas), representación simbólica, conducta intencional [13].

En este punto, no obstante, la comunicación —que podemos denominar proto-lingüística— tiene lugar sin ningún tipo de especialización cognitiva, sin que los participantes en el proceso compartan ningún tipo de conocimiento específicamente lingüístico. Se trata, más bien, de una comunicación basada en la capacidad de comportarse intencionalmente, y de comprender acontecimientos, gracias al efecto clarificador del contexto (surgiría en este punto, por ejemplo, la ostensión, el uso indicativo del dedo índice). Se trata, de hecho, de una capacidad que, como hemos visto, no presupone mucho más que lo que algunos chimpancés bien entrenados han conseguido alcanzar, con excepción de una rudimentaria habilidad de articulación vocal, y de una mayor capacidad cognitiva, que permite suponer una mayor capacidad de combinar signos y comprender series de asociaciones y de actuar intencionalmente. Nuestra hipótesis con respecto al lenguaje es que, llegados a este punto, esta capacidad elemental de comunicación, por su significación adaptativa, es seleccionada, lo que conlleva la aparición de estructuras neurales especializadas en la producción y comprensión de estas vocalizaciones, a partir de estructuras preexistentes (como la especializada en reconocer llamadas vocales de miembros de la propia especie, o en la coordinación motora).

Este proceso es explicable en razón de mecanismos evolutivos generales. La idea es que ciertas condiciones evolutivas favorecen el desarrollo de mecanismos innatos especializados, cuyo funcionamiento es inconsciente, automático y sin esfuerzo, que Fodor (1983) llama módulos. Estas condiciones se refieren a aquellos aspectos del ambiente de importancia vital para el organismo, siempre y cuando sean robustos, esto es, no cambien a un ritmo más rápido de la capacidad del genoma para ser sensible a tales cambios (Plotkin y Odling-Smee, 1979). Por ejemplo, nos resulta vital detectar precipicios, calcular trayectorias de objetos en movimiento o reconocer rostros. Igualmente, ser capaz de acceder rápidamente al lenguaje del propio grupo; dado su papel clave, es vital para el desarrollo infantil. La existencia de comunicación aún proto-lingüística crea una presión selectiva en favor de que este acceso sea facilitado innatamente, para poder acceder a ella cuanto antes. De esta forma, lo aprendido en una generación puede influir y acelerar el proceso evolutivo. El mecanismo por el que lo que se aprende en una generación influye en la condición genética de la siguiente, imitando así de forma superficial el mecanismo lamarckiano, es conocido como «efecto Baldwin» (Hinton y Nowlan, 1987). Además, este efecto conlleva que, a medida que au-

13. En el sentido de conducta propositiva; no implica, por tanto, la capacidad para la conciencia reflexiva, para la atribución de intenciones complejas. Estos aspectos serían, más bien, una consecuencia de la adquisición del lenguaje.

menta la especialización innata, la presión que convierte en adaptativo el proceso decae, dando lugar a un punto de equilibrio entre lo innato y lo aprendido, pudiendo alterar a su vez la naturaleza de esta nueva capacidad.

Este modelo se adecua bien al caso del lenguaje. La presión por facilitar la adquisición de las claves para acceder a las prácticas comunicativas pudo llevar a desarrollar una progresiva especialización neuronal, pero no hasta el punto de preprogramar todo el lenguaje, dada la constante renovación del proto-lenguaje, para mejorar su poder expresivo y su eficiencia. Igualmente, este modelo encaja con el necesario proceso de desarrollo cerebral en la infancia, derivado como vimos de su aumento de tamaño relativo. El desarrollo del cerebro constituye un caso ejemplar de equilibrio entre especialización y flexibilidad, de desarrollo genéticamente controlado y desarrollo sensible a las contingencias ambientales (Changeux, 1985). La existencia de un periodo crítico para la adquisición del lenguaje constituye la mejor evidencia de este proceso de especialización innata y flexibilidad.

La aparición de un mecanismo innato dedicado al lenguaje [14]: como nos indica la Lingüística histórica, los cambios sintácticos están presentes en todas las lenguas, a un ritmo que no puede ser seguido por los cambios genéticos (Wang, 1976). La culminación del proceso, el paso definitivo al lenguaje, lo marcaría la aparición de la capacidad recursiva, la clave, como vimos, de la distintiva productividad del lenguaje humano.

En resumen, las características actuales del lenguaje constituyen el resultado (siempre provisional, como todo lo sometido a la evolución) de la interacción entre el ejercicio de la inteligencia general, consecuencia a su vez de un mayor cerebro, y la selección a nivel genético de ciertas predisposiciones que facilitan los aspectos más vitales de este ejercicio, favoreciendo un equilibrio entre lo genéticamente programado y lo que hay que aprender, siempre presente por la posibilidad de innovar que permite este aumento cerebral.

14. Tanto en su aspecto productivo como comprensivo. Como ha mostrado Hurford (1989), la única estrategia evolutivamente estable para la fijación de este módulo especializado es utilizar los mismos recursos en la producción y la comprensión. Ello conlleva su independencia de modalidad sensorial, lo que permite explicar la posibilidad de lenguajes no vocales (como el de señas, por ejemplo). Produciría un efecto facilitador en su adquisición y uso, sin eliminar por ello la intervención de los mecanismos cognitivos generales que permitieron inicialmente el surgimiento de comunicación vocal. Se entraría así en un proceso de aumento gradual de la complejidad gramatical, en función de las necesidades comunicativas encontradas y los recursos disponibles. Bickerton (1990) rechaza la idea de cambio gradual desde el proto-lenguaje al lenguaje, en favor de la explicación por medio de un salto evolutivo. Sin embargo, abundan los ejemplos (pidgins, lenguaje de turistas o emigrantes, aprendices de una segunda lengua...) que ilustran la posibilidad de un continuo de sistemas de comunicación de diversa complejidad gramatical.

VI. CONCLUSIÓN

En este trabajo hemos intentado reconstruir el proceso evolutivo por el que apareció el lenguaje humano. Nuestra hipótesis —que el lenguaje surgió para satisfacer las necesidades comunicativas de nuestros antepasados— pretende hacer inteligible la aparición del lenguaje en el marco del surgimiento de nuestra especie, al ligarla al proceso de encefalización y a la adopción de la estrategia cognitiva como vía adaptativa, la estrategia de la novedad, de la flexibilidad adaptativa. En este punto, la pregunta que nos planteábamos anteriormente con respecto a la posibilidad de un lenguaje totalmente innato puede ser respondida con confianza negativamente: el sentido evolutivo del lenguaje responde a las necesidades comunicativas de organismos cognitivos, lo que supone la posibilidad de infinitos mensajes (frente a la limitación del repertorio de los monos vervets, por ejemplo). La forma en que el lenguaje satisface su función, mediante una jerarquía de niveles de organización gramatical, por su sistematicidad, permite asegurar los medios de formular esos contenidos ilimitados por medio de medios finitos, permitiendo de este modo que puedan ser entendidos. Para ello, es preciso atender a las prácticas comunicativas de la propia comunidad, a las que uno, no obstante, se encuentra ya predispuesto por razones asimismo evolutivas. La colonización de toda la Tierra por parte de nuestra especie constituye la mejor prueba del éxito de esta estrategia.

Nuestro escenario evolutivo evita, por otra parte, las paradojas que suelen afectar con frecuencia a las hipótesis sobre el origen del lenguaje, bien porque no se tiene en cuenta más que un aspecto del proceso, bien porque se considera este origen al margen de otros aspectos conectados con el propio origen del lenguaje. Por nuestra parte, hemos propuesto diversas etapas en el proceso, distinguiendo en cada una de ellas lo que es seleccionado de las consecuencias que se siguen a continuación, dando lugar a nuevos desarrollos, y distinguiendo igualmente entre las presiones selectivas en acción de los medios disponibles para afrontarlas. Así, hemos explicado la aparición de la capacidad articulatoria de la presión por un medio de comunicación, y hemos derivado la justificación de un enfoque innatista para la ontogenia del lenguaje de un enfoque funcionalista de su filogenia. Igualmente, la capacidad intencional que hemos postulado como base de la comunicación proto-lingüística no implica conciencia reflexiva, ni por tanto la idea de que el lenguaje resulta de una convención.

La conciencia reflexiva, por el contrario, constituiría más bien una consecuencia de la propia evolución del lenguaje, al igual que el sentido de identidad personal. Y, sin duda, el impacto del lenguaje en la propia forma de vida humana debió de ser enorme. En este sentido, quisiera señalar solamente el que puede ser considerado punto culminante del pro-

ceso: la aparición del *Homo Sapiens Sapiens*, hace unos 35.000 años, marcada por una espectacular explosión tecnológica (con la novedad de la fabricación de herramientas destinadas a la fabricación de herramientas), los primeros enterramientos rituales, pinturas y dibujos en cuevas y objetos para adorno personal (Noble y Davidson, 1991), signos inequívocos de la complejidad simbólica y cultural alcanzada gracias al lenguaje.

La tarea, sin embargo, dista todavía de estar concluida. Hemos dejado de lado la cuestión —más compleja, más difícil— de la periodización de este proceso dentro de la familia Homínida. Sería preciso considerar las aportaciones de la Antropología física y la ecológica para refinar el cómo y el cuándo del proceso evolutivo cuya lógica hemos bosquejado. Es una tarea, sin duda, que excede mis capacidades.

BIBLIOGRAFÍA

Akmajian, A., Demers, R. y Harnish, R. (1984), *Linguistics: an introduction to language and communication*, MIT Press. V.e.: Alianza, Madrid, 1992.
Albrecht, G. A. (1978), «The craniophacial morphology of Sulawesi macaques»: *Primatology*, 3.
Bates, E. (1976), *Language and context: studies in the acquisition of pragmatics*, Academic Press.
Bates, E., Thal, D. y Marchaman, V. (1989), «Symbols and syntax: a Darwinian approach to language development», en N. Krasnegor, *et al.* (eds.), *The Biological Foundations of Language Development*, Oxford University Press.
Bennett, J. (1976), *Linguistic Behavior*, Cambridge University Press.
Bickerton, D. (1984), «Lenguas criollas»: *Investigación y Ciencia*, julio.
Bickerton, D. (1990), *Language and Species*, University of Chicago Press.
Bickhard, M. H. (1979), «On Necessary and Specific Capabilities in Evolution and Development»: *Human Development*, 22, 217-224.
Biggerstaff, R. H. (1977), «The biology of the human chin», en A. A. Dahlberg y T. M. Graber (eds.), *Orofacial growth and development*, Mouton, Den Hagen, 71-87.
Bradshaw, J. L. y Nettleton, N. C. (1989), «Lateral Asymmetries in Human Evolution»: *The International Journal of Comparative Psychology*, 3, 37-71.
Brandon, R. N. y Hornstein, N. (1986), «From Icons to Symbols: some speculations on the origins of language»: *Biology and Philosophy*, 1, 169-189.
Bruner, J. (1983), *Child's Talk: Learning how to use Language*, Norton, New York.
Changeux, J. P. (1985), *Neuronal Man: The Biology of Mind*, Pantheon, New York.
Cheney, D. L. y Seyfarth, R. M. (1990), *How monkeys see the world: inside the mind of another species*, University of Chicago Press, Chicago.
Chomsky, N. (1980), *Rules and Representations*, Blackwell. V.e.: FCE, 1986.
Chomsky, N. (1981), *Lectures on Government and Binding*, Foris.
Chomsky, N. (1988), *Language and problems of knowledge: The Managua lectures*, MIT. V.e.: Visor, Madrid, 1992.

Clutton-Brock, T. H. y Harvey, P. H. (1980), «Primates, Brains and Ecology»: *Journal of Zoology*, *190*, 309-323.
Condillac, E. B. (1746), *Essai sur l'origine des connaissances humaines, ouvrage ou l'on reduit à un seul principe tout ce que concerne l'entendement.*
Corballis, M.C. (1992), «On the evolution of language and generativity»: *Cognition*, *44*, 197-226.
Davidson, D. (1982), «Rational Animals»: *Dialectica*, *36*, 317-327.
Dawkins, R. (1986), *The Blind Watchmaker*, Longman, London. V.e.: Labor, Barcelona, 1988.
Deacon, T. W. (1988), «Human Brain Evolution, I & II», en H. J. Jerison e I. Jerison (eds.), *Intelligence and Evolutionary Biology*, Springer-Verlag, Berlin, 363-415.
Dunbar, R. I. M. (1993), «Coevolution of neocortical size, group size and language in humans»: *Behavioral and Brain Sciences*, *13*.
Falk, D. (1975), «Comparative anatomy of the larynx in man and the chimpanzee: implications for language in Neanderthal»: *American Journal of Physical Anthropology*, *43*, 123-132.
Fodor, J. A. (1983), *The modularity of mind*, MIT Press, Cambridge, Mass.
Foley, R. (1987), *Another unique species: patterns in human evolutionary ecology*, Longman.
Foley, R. (1990), «The causes of brain enlargement in human evolution»: *Behavioral and Brain Sciences*, *13*, 354-356.
García-Albea, J. E. (1993), «La capacidad humana del lenguaje: un ejemplo de discontinuidad evolutiva», en Id., *Mente y Conducta*, Trotta, Madrid.
Gardner, R. A. y Gardener, B. T. (1969), «Teaching sign language to a chimpanzee»: *Science*, *165*, 664-672.
Gazzaniga, M. S. (1983), «Right hemisphere language following brain bisection»: *American Psychologist*, *3*, 525-537.
Gómez, J. C., Sarria, E. y Tamarit, J. (1993), «The comparative study of early communication and theories of mind: ontogeny, phylogeny and pathology», en S. Baron-Cohen *et al.* (eds.), *Understanding other minds: perspectives from autism*, Oxford University Press, 397-426.
Gomila, A. (1992), «Acerca de la naturaleza psicológica de la Lingüística»: *Theoria*, *16-17-18*, 973-987.
Gomila, A. (en prensa), «The functionality of the study of language origin»: *Behavioral and Brain Sciences* (continuing commentary on Pinker & Bloom, 1990).
Gould, S. J. (1987), «Integrity and Mr. Rifkin», en S. J. Gould (ed.), *An Urchin in the storm*, Norton, New York, 195-249.
Gould, S. J. y Lewontin, R. C. (1979), «The spandrels of San Marco and the Panglossian programme: a critique of the adaptationist programme»: *Proceedings of the Royal Society of London*, *205*, 281-288.
Gould, S. J. y Vrba, E. S. (1982), «Exaptation: a missing term in the science of form»: *Paleobiology*, *8*, 4-15.
Greenfield, P. y Smith, J. (1976), *The structure of communication in early language development*, Academic Press, New York.
Grimshaw, J. y Pinker, S. (1989), «Positive and negative evidence in language acquisition»: *Behavioral and Brain Sciences*, *12*, 341-342.

Hailman, J. P. y Ficken, M. S. (1986), «Combinatorial animal communication with computable syntax: Chick-a-dee calling qualifies as "language" by structural linguistics»: *Animal Behaviour*, *34*, 1899-1901.
Harnad, S., Steklis, H.D. y Lancaster, J. (eds.) (1976), «Origins and Evolution of Language and Speech»: *Annals of the New York Academy of Sciences*, vol. 280.
Herman, L. M., Richards, D. G. y Wolz, J. P. (1984), «Comprehension of sentences by bottle-nosed dolphins»: *Cognition*, *16*, 129-219.
Hewes, G. W. (1973), «Primate communication and the gestural origins of language»: *Current Anthropology*, *14*, 5-24.
Hinton, G. E. y Nowlan, S. J. (1987), «How learning can guide evolution»: *Complex Systems*, *1*, 495-502.
Hurford, J. (1987), «Language and Number. The emergence of a cognitive system», Blackwell.
Hurford, J. (1989), «Biological evolution of the Saussuren sign as a component of the language acquisition device»: *Lingua*, *77*, 187-222.
Jacob, F. (1977), «Evolution and tinkering»: *Science*, *196*, 1161-1166.
Jerison, H. J. (1976), «Paleoneurology and the evolution of mind»: *Scientific American*, *234*, 90-101.
Klima, E. y Bellugi, U. (1979), *The signs of language*, Harvard University Press.
Krebs, J. R. y Dawikins, R. (1984), «Animal signals: mind-reading and manipulation», en J. R. Krebs y N. B. Davies, (eds.), *Behavioural Ecology: an evolutionary approach*, Blackwell.
Krogman, W. M. (1972), *Child growth*, University of Michigan Press, Ann Arbor.
Lewis, D. (1969), *Convention*, Harvard University Press.
Lieberman, P. (1984), *The Biology and Evolution of Language*, Harvard UP.
Lightfoot, D. (1989), «The child's trigger experience: degree-0 learnability»: *Behavioral and Brain Sciences*, *12*, 321-334.
Lorenz, K. (1977), *Behind the Mirror: a search for a natural history of human knowledge*, Methuen, London.
Marcus, G. F. (1993), «Negative Evidence in Language Acquisition»: *Cognition*, *46*, 53-85.
Mehler, J. (1985), «Review of P. Lieberman's "The Biology and Evolution of Language"»: «*Journal of the Acoustic Society of America*, *80*, 1558-1560.
Mehler, J. y Dupoux, E. (1990), *Naitre Humain*, Odile Jacob, Paris. V.e.: Alianza, Madrid, 1992.
Milton, K. (1993), «Dieta y nutrición de los primates»: *Investigación y Ciencia*, *205*, 56-63.
Morgan, J. L. (1989), «Learnability considerations and the nature of trigger experiences in language acquisition»: *Behavioral and Brain Sciences*, *12*, 352-353.
Newmeyer, F. J. (1991), «Functional explanations in Linguistics and the origins of language»: *Language and Communication*, *11*, 3-28.
Newport, E. L. (1990), «Maturational constraints on language learning»: *Cognitive Science*, *14*, 11-28.
Noble, W. y Davidson, I. (1991), «The evolutionary emergence of modern human behaviour: language and its archeology»: *Man*, *26*, 223-253.
Piatelli-Palmarini, M. (1989), «Evolution, selection and cognition: from "lear-

ning" to parameter-setting in biology and the study of language»: *Cognition*, *31*, 1-44.

Pinker, S. y Bloom, P. (1990), «Natural Language and Natural Selection»: *Behavioral and Brain Sciences*, *13*, 707-727.

Plotkin, H. C. y Odling-See, F. J. (1979), «Learning, Change and Evolution: an enquiry into the teleonomy of learning»: *Advances in the Study of Behavior*, *10*, 1-39.

Potter, B. (1986), «The allometry of primate skeletal weight»: *International Journal of Primatology*, *7*, 457-466.

Premack, D. (1985), «Gavagai! or the future history of the animal language controversy»: *Cognition*, *19*, 207-296.

Premack, D. (1988), «Minds with and without language», en L. Weiskrantz, (ed.), *Thought without language*, Clarendon, 46-65.

Rak, Y. (1987), «What's so robust about the hyper-robust Australopithecus boisei?»: *American Journal of Physical Anthropology*, *72*, 244.

Sacks, O. (1989), *Seeing Voices*, University of California Press. V.e.: Anaya y Mario Muchnik, 1991.

Savage-Rumbaugh, E. (1986), *Ape language: from conditioned response to symbol*, Oxford University Press.

Savage-Rumbaugh, E. (1990), «Language as a cause-effect communication system»: *Philosophical Psychology*, *3*, 55-76.

Savage-Rumbaugh, E., McDonald, K., Sevcik, R. A., Hopkins, W. D. y Rupert, E. (1986), «Spontaneous symbol acquisition and communicative use by pygmy chimpanzee (Pan Paniscus)»: *Journal of Experimental Psychology: General*, *112*, 211-235.

Searle, J. (1979), «Intentionality and the Use of language», en A. Margalit, (ed.), *Meaning and Use*, D. Reidel, Derdrecht, 181-197.

Seidengerg, M. S. y Petitto, L. A. (1987), «Communication, symbolic communication, and language: comment on Savage-Rumbaugh *et al.* (1986)»: *Journal of Experimental Psychology: General*, *116*, 279-287.

Smith, C. G. (1985), *Ancestral Voices*, Prentice-Hall.

Snowdon, C. T. (1982), «Linguistic and Psycholinguistic approaches to primate communication», en C. T. Snowdon, C. H. Brown y M. R. Petersen (eds.), *Primate Communication*, Cambridge University Press, 212-238.

Talmy, L. (1988), «Force dynamics in language and cognition»: *Cognitive Science*, *12*, 49-100.

Terrace, H. S. (1979), «Is problem solving language?»: *Journal of the Experimental Analysis of Behavior*, *31*, 161-175.

Tennant, N. (1984), «Intentionality, syntactic structure, and the evolution of language», en C. Hookway, (ed.), *Minds, Machines and Evolution*, Cambridge University Press, 73-103.

Tobias, P. V. (1987), «The brain of Homo Habilis: A new level of organization in cerebral evolution»: *Journal of Human Evolution*, *16*, 741-761.

Wang, W. S. Y. (1976), «Language change», en Harnad *et al.* (eds.), 1976.

Wells, G. A. (1987), *The Origin of Language. Aspects of the discussion from Condillac to Wundt*, Open Court, La Salle, Illinois.

Wilkins, W. y Dumford, J. (1990), «In defense of exaptation»: *Behavioral and Brain Sciences*, *13*, 763-764.

Wilson, E. O. (1975), *Sociobiology: the New Synthesis*, Harvard University Press. V.e.: Omega, 1980.

Wimmer, H. y Perner, J. (1983), «Beliefs about beliefs: Representations and constraining functions of wrong beliefs in young children's understanding of deception»: *Cognition*, *18*, 103-128.

EL CONTROL RACIONAL DE LA CONDUCTA

Fernando Broncano

«Say from whence you owe this strange intelligence?»
Shakespeare

I. LAS DIMENSIONES DE LA RACIONALIDAD

La racionalidad consiste en el uso teórico, práctico y evaluativo de la razón. En otras palabras, en el uso de la razón para adoptar creencias, tomar decisiones y evaluar hechos. La razón es la facultad o capacidad para pensar y actuar inteligentemente. La inteligencia es, a su vez, la capacidad para encontrar, plantear y resolver problemas. Esta es una definición de racionalidad (Rescher, 1993) entre otras varias que nos ofrecen los filósofos y psicólogos, pero tiene la virtud de recoger los tres aspectos más importantes de la racionalidad, el razonamiento que conduce a la fijación o aceptación de creencias, el razonamiento que conduce a la acción y, por último, el razonamiento que nos permite elegir fines y valorar hechos. Cada uno de estos aspectos recibe un tratamiento diferenciado en la epistemología, en la teoría de la acción y en la ética; el nuestro, desde el punto de vista de la filosofía de la mente, se interesa por el tema general del origen y explicación de esta facultad que llamamos razón.

Desde este punto de vista, el primer problema filosófico que encontramos es el de explicar cómo fueron posibles las razones en un mundo de causas y qué tipo de extraña función es la razón dentro de nuestro sistema biológico de funciones. La racionalidad parece ser un rasgo distintivo de nuestra especie. Está ligada al hecho de que tenemos representaciones y que obtenemos nuevas representaciones mediante inferencias. Desde un punto de vista biológico, la racionalidad ha evolucionado

como un sistema de control de inferencias mediante normas. Llamamos normas de racionalidad a las reglas que nos permiten producir inferencias aceptables o adecuadas. Actúan como meta-representaciones, que tienen relación con el contenido que regulan, aunque pueden automatizarse en esquemas formales de inferencia.

Si creyésemos en la historia de A. Clarke, que nos muestra Stanley Kubrick en «2001, Una odisea en el espacio», según la cual la inteligencia fue inducida en los primates por otros organismos inteligentes, no nos preocuparía cómo la inteligencia pudo emerger desde sistemas que no son en sí mismos inteligentes, y quizás nos sintiésemos complacidos con una mera descripción fenomenológica de lo que llamamos racionalidad. Pero si creemos, como es racional hacerlo, en la aparición evolutiva del control racional de la conducta, la descripción fenomenológica que obtenemos por introspección no será suficiente. O no lo será más de lo que es la introspección en el estudio científico de la conducta: una parte importante, esencial en ocasiones, pero una parte de un movimiento más amplio que conjuga la introspección con la observación externa o heterofenomenología (Dennett, 1992, c 4).

En segundo lugar, la racionalidad tiene una dimensión esencialmente cognitiva, es un mecanismo o función cognitiva que se ocupa de la manipulación de representaciones y que, cuando tiene ciertas virtudes, es considerada racional. La racionalidad es, pues, una virtud cognitiva, una facultad que debe ser ejercida adecuadamente, que debe producir buenos, aceptables o exitosos resultados. En otro caso no es llamada racionalidad. Los locos manipulan sus representaciones, pero no lo hacen racionalmente. No hay racionalidad sin representaciones y no hay representaciones sin contenido, de ahí que la racionalidad se nos aparezca bajo una forma bifronte, de un lado desde la dimensión subjetiva del sujeto y del otro desde la posición objetiva del observador externo. Una de las principales fuentes de problemas filosóficos deriva de esta doble cara de la racionalidad. La dimensión subjetiva es necesaria para limitar la racionalidad a la perspectiva del sujeto, pues algo es racional no de manera intrínseca, sino relativamente, entre otras cosas, a las luces y a la perspectiva del sujeto. Más, por la misma razón, la perspectiva subjetiva no es ni puede ser suficiente, no es la autoridad última que decida si algo es o no racional. El sujeto no puede ser el juez y la parte. El loco piensa de sí mismo que está actuando racionalmente, pero no lo está, objetivamente hablando. Por eso necesitamos la dimensión externa, social y medioambiental. Ser racional es algo así como dominar o ser maestro en ciertas habilidades, una maestría que sólo pueden sancionar el resto de los sujetos racionales. Y ser racional es serlo en relación a ciertas circunstancias objetivas que nos son independientes. Desde el punto de vista científico y filosófico, ésta es la dimensión por la que debemos comenzar, porque la perspectiva subjetiva nos resulta oculta, teórica, y

solamente la inferimos analizando el comportamiento del sujeto en circunstancias que nos son reconocibles.

En tercer lugar, la racionalidad es nuestro principal instrumento de supervivencia, pero, sobre todo, es el cemento de la sociedad. La cooperación social, esporádica en la acción particular o permanente en las instituciones, exige profundos lazos y mecanismos afectivos y emotivos, que serían inútiles si no fuéramos capaces de entender y predecir la conducta de los otros (Dennet, 1978, 1987; Bennett, 1991). La cooperación y la competencia, el egoísmo y el altruismo, el amor y la guerra exigen la facultad de predecir y adelantarse al otro. Sin esa facultad seríamos incapaces siquiera de bajar una mesa entre dos personas por la escalera de nuestro piso. Esta exigencia pone en primer plano el carácter adscriptivo de la racionalidad, el punto de vista externo que se dirige a la subjetividad de otros para cumplimentar los propios deseos. De manera que no abandonamos la noción subjetiva por un prurito conductista, sino por una más compleja concepción del yo en el nosotros y el nosotros, a su vez, en un medio ecológico, espacio-temporal.

La racionalidad, por último, es un término esencialmente normativo, que alude al aprovechamiento máximo de los medios para alcanzar los objetivos en el menor tiempo posible, con el menor gasto de recursos. En un mundo en el que la información es limitada y difícil de conseguir, en el que nuestros recursos son pocos y nuestro tiempo corto, la normatividad de la racionalidad nos exige buscar los medios óptimos y hacer los mínimos gastos para conseguir los máximos objetivos que podamos. Esta es la característica más sobresaliente de la racionalidad. Pero esta exigencia optimizadora es, a su vez, una nueva fuente de problemas porque, lo mismo que una ética utópica suele generar comportamientos morales hipócritas y filisteos, una insistencia en el aspecto ideal de la racionalidad convierte a los sujetos empíricos en animales esencialmente irracionales.

II. LA ATRIBUCIÓN DE CREENCIAS Y DESEOS: EL MODELO CANÓNICO DE RACIONALIDAD

Se atribuye racionalidad a otros seres para explicar su conducta (Dennett, 1987). La racionalidad es la marca de fábrica de la intencionalidad. Si atribuimos racionalidad a la conducta ajena, su conducta queda calificada por ello como conducta intencional, y, según algunos autores, también a la inversa [1]. Un carraspeo puede significar un resfriado, pero

1. «Puesto que la identidad de un pensamiento no puede divorciarse de su lugar en la red lógica de otros pensamientos, los pensamientos, como las proposiciones, no se pueden resituar en la red sin convertirse en pensamientos diferentes. La incoherencia radical en la creencia es imposible. Tener una actitud proposicional solitaria es tener una lógica correcta en su mayor parte» Davidson (1982, 475).

también una señal comunicativa intencional. Al conceder intencionalidad a tal conducta, el hecho neurofisiológico del carraspeo se convierte en un hecho social de comunicación que nos exige otro tipo de explicación. Durante muchos años se ha discutido el tipo de explicación adecuada a las ciencias humanas, si es irremisiblemente intencional o si debe ser causal en último extremo, pero en la vida cotidiana la explicación no está regida por las consideraciones metodológicas de los filósofos, sino por la exigencia perentoria de predecir la conducta ajena en el momento adecuado. El no hacerlo puede significar la exclusión de la comunidad, cuando no está en cuestión la supervivencia misma. Porque la predicción de la conducta ajena es también la condición para que el otro forme parte del nosotros. Y probablemente este hecho fue determinante en la emergencia de la racionalidad, en un cerebro cuya principal función es anticiparse al curso de la naturaleza para ordenar la respuesta correcta.

En el dominio de la acción intencional hay una perfecta simetría entre la explicación y la predicción. Esta simetría deriva de lo que llamaremos el *modelo «deseos-creencias»* de explicación de la acción: explicamos la acción A del sujeto X en las circunstancias C cuando atribuimos a X una razón para A. Simétricamente, predecimos que X hará A en C si conjeturamos que tiene una razón para hacerlo. A. Goldman (1971) y D. Davidson (1963) han sido los principales promotores y defensores de este modelo, que se resume en el siguiente esquema de explicación:

1) X desea O en C
2) X cree que haciendo A consigue O
3) X hace A

(1) y (2) deben ser causas suficientes para (3). Aunque es una explicación en la que no están involucradas leyes, sí es, según Davidson, una explicación causal de la conducta. Concedemos el nombre de deseo a una pro-actitud que abarca un gran número de estados mentales de carácter motivacional incluyendo, junto a lo que solemos considerar normalmente como deseos, toda la lista de estados motivacionales que involucren la búsqueda de un objetivo. Así, el miedo a las arañas es motivacional cuando conlleva el objetivo de alejarse del peligro. Lo mismo ocurre con lo que llamamos creencia: resume todos los estados mentales en los que el agente da por buena cierta información.

Este modelo está expuesto en el lenguaje de la psicología natural, que es nuestra dotación común para conseguir y mantener la convivencia o supervivencia social. Pese a que puede ser atacado de poco científico (véase J. Toribio, «El eliminativismo y la Psicología Popular» en este vo-

lumen), puede expresarse en los términos mucho más refinados de la teoría de la decisión. En este caso es necesario que el sistema deseos-creencias cumpla ciertos requisitos formales de los que se ocupa la teoría formal de la decisión. Para comenzar, la exposición cualitativa del deseo sería sustituida por un valor numérico que indicaría la utilidad que un individuo concede a un objetivo o estado de cosas producido por su acción. Del mismo modo, la creencia estaría representada por una asignación de probabilidad a un estado de la naturaleza del que depende la consecución del objetivo útil, que representaría el conocimiento que tiene el sujeto del mundo, en virtud del cual asigna grados de expectación a los sucesos. Necesitamos, además, un conjunto de cursos de acción o espacio de decisiones que están a disposición del agente y de quien presuponemos que es libre y capaz de optar por cada uno de ellos. Un cuarto supuesto necesario para que podamos usar la teoría de la probabilidad es la independencia de las probabilidades de cada estado respecto a las alternativas elegidas. Este supuesto, traducido al lenguaje de deseos-creencias, establece que ambos deben ser independientes: ni los deseos deben modificar causalmente nuestras expectativas de sucesos, ni, a la inversa, nuestro conocimiento de las probabilidades debería modificar nuestros deseos.

Al atribuir racionalidad a un agente le estamos atribuyendo *a*) la capacidad de representarse un conjunto de estado posibles del mundo, *b*) un conjunto de objetivos o resultados, *c*) un conjunto de cursos de acción posibles y *d*) una función de elección que tiende a conseguir el máximo de satisfacción de esos objetivos. Supongamos, para emplear el ejemplo tradicional (llamado «la tortilla de Savage» por ser el que empleó Savage en su análisis de la decisión en contextos de incertidumbre) que al llegar a casa con hambre observamos que nuestra compañera o compañero ha comenzado a hacer una tortilla y ha cascado 5 de los 6 huevos que llevaban ya mucho tiempo en la nevera. Puede que el sexto esté podrido, así que debemos tomar una decisión sobre echarlo al recipiente, dejarlo donde está u observar su estado en una taza aparte. La matriz que representaría nuestras posibilidades en el modelo deseos-creencias es (*véase página siguiente*).

Según este esquema, podríamos predecir la conducta de una persona normal, puesto que el curso de acción más favorable, y al que corresponde la decisión más racional, es el segundo: no perdemos nada por comprobar el estado del huevo y en el mejor de los casos ganamos una tortilla mayor. La matriz anterior es algo así como un modelo del cerebro de otra persona, que nos permite explicar su decisión, o predecirla si es el caso. Los deseos están representados por los objetivos y las creencias por su conocimiento de cómo son o cómo pueden ser las cosas y cómo afecta a sus deseos el tomar un curso u otro de acción.

Este modelo se ha criticado por muchas razones y desde múltiples

Cursos de acción	Estados del mundo		
	podrido	*buen estado*	
cascarlo en el bol	no hay tortilla	tortilla de seis huevos	Resultados obtenidos
tirar el huevo	tortilla de cinco huevos	tortilla de cinco huevos	
comprobar en un vaso aparte	tortilla de cinco huevos	tortilla de seis huevos	

perspectivas, mas reparemos antes de nada en sus ventajas[2]. Es simple y, no obstante, permite expresar con precisión las condiciones de racionalidad, pues aun bajo la forma matemática de la teoría de la decisión se puede comprender fácilmente cómo el modelo conforma una teoría de la racionalidad. Una teoría descriptiva, en primer lugar, dado que el modelo sirve para anticiparnos a la conducta de los otros. Si no hubiera tenido éxito en este sentido, no estaría tan atrincherado en nuestro esquema conceptual y habría desaparecido como la medicina popular o la física popular han desaparecido de nuestra concepción racional del mundo. Mas, por otra parte, configura una teoría normativa, ya que establece que, en principio, el agente debe escoger y escogerá la opción que suponga para él un máximo beneficio con el mínimo coste o riesgo. Esta perspectiva normativa es la que hace del modelo deseos-creencias propiamente una teoría de la racionalidad, puesto que la normatividad que se postula presupone el cumplimiento de ciertos requisitos que coinciden con los que tradicionalmente consideramos la marca de fábrica de la racionalidad. Estos requisitos se resumen en los siguientes:

2. La crítica del modelo económico se ha convertido ya en un género literario y posiblemente en una profesión académica. De entre las infinitas ofertas en el mercedo, Hollis (1987), Parfit (1984) y, sobre todo, la recopilación de Sen y Williams (1982) permiten una perspectiva amplia, inteligente y matizada. Entre nosotros, Muguerza (1977 y 1990) forman ya parte del patrimonio genético-cultural de mi generación.

a) La coherencia en los conocimientos de los que dispone el sujeto

Esta breve enunciación esconde un principio fuerte, discutible y discutido: la racionalidad práctica presume la racionalidad teórica, que se fundamenta en la coherencia de los contenidos. El preferidor ideal es también un individuo lógico y racional en lo epistémico. En la versión más simple del modelo se presume que su grado de racionalidad es máximo, o lo que es lo mismo, sin restricciones a su capacidad de cálculo lógico (Davidson, 1982; Dennet, 1987 y Cohen, 1981).

b) La compleción y coherencia entre los objetivos y el orden de las preferencias

Entre los varios principios que configuran el principio de orden de las preferencias, está el que cada dos alternativas son, o bien indiferentes, o bien preferibles la una a la otra. Este principio ha recibido numerosas críticas por su idealización. De entre ellas, la más interesante es la de Putnam, 1988, quien señala que, además de indiferentes, dos alternativas podrían ser incomparables, como pudiera ser el caso de la famosa apuesta pascaliana entre la vida placentera y la condenación eterna. La crítica de Putnam pertenece a una familia de críticas de principio a la teoría de la decisión por su incompleción y por la necesidad de ser complementada por principios éticos. Otra familia de críticas son internas y se dirigen al principio de Savage conocido como «Principio de la cosa segura», que establece que dos alternativas de acción deben ser ordenadas solamente por sus diferencias en los valores que aparecen en las filas y columnas, no atendiendo a aquellos valores comunes que aparezcan en más de una, lo que entraña que, en muchas ocasiones, una minúscula diferencia en el valor de un resultado ordene contraintuitivamente cursos de acción que deberían ser indiferentes, como, por ejemplo, la alternativa de tío Gilito el millonario de bajarse o no del Rolls Royce para recoger una peseta. Dentro de este tipo de críticas internas está la llamada por Robert Nozick «Paradoja de Newcomb», que se presenta cuando consideramos que en un orden determinístico de la naturaleza los estados del mundo ya están prefijados, por lo que nuestro razonamiento puede hacer variar nuestra elección dependiendo de si atendemos a que las cosas pueden ocurrir con cierta probabilidad o a que ya han sido determinadas (Nozick, 1970). Puesto que nuestro objetivo es considerar la noción de racionalidad desde un punto de vista naturalista y no evaluar la adecuación de la teoría de la preferencia racional, no consideraremos estas críticas más allá de esta nota marginal, que no entraña, claro está, que las críticas no estén bien fundadas.

c) La maximización de los beneficios en la opción escogida

Este principio de maximización se extiende a la utilidad condicional en el caso de que los estados del mundo no sean independientes y se establezcan relaciones probabilísticas entre sus ocurrencias (Jeffrey, 1965; Gardenförs y Sahlin, 1988, 9). Al igual que en el caso anterior, tampoco está muy claro qué significa maximizar los beneficios y minimizar los costos. Tradicionalmente se ha pretendido precisarlo mediante dos principios o estrategias racionales que atienden a una mayor insistencia en la utilidad o, por el contrario, en el riesgo asumido al optar por una u otra alternativa. El llamado *principio maximin* elige la alternativa que hace máximo el beneficio más pequeño. Es la estrategia más conservadora desde el punto de vista del riesgo. La estrategia de *minimax* elige la alternativa que hace menor el riesgo máximo de pérdidas. Los teóricos de las preferencias han discutido mucho las ventajas e inconvenientes de uno y otro, así como de otros derivados, entre los que se encuentra el llamado «principio de razón insuficiente» (Luce y Raiffa, 1957). También aquí señalamos solamente la dirección para seguir investigando, puesto que tales discusiones solamente afectan marginalmente a nuestra discusión.

La teoría de la decisión se formula en lenguaje matemático y postula requisitos que garantizan que un razonador ideal esté a salvo de inconsistencias que surjan de la propia teoría. Bien es cierto que las distintas formulaciones y los requisitos para que el agente cumpla los criterios de racionalidad han consumido una montaña de trabajos por parte de algunos de los mejores matemáticos y economistas de nuestro siglo (Ramsey, 1931; Savage, 1954; Luce y Raiffa, 1957). Pese a todo, tanto Ramsey como Savage probaron que si el agente obedece ciertos requisitos formales (que resumen los criterios ya relatados), es posible adscribirle unívocamente una función interna o subjetiva de probabilidad y una medida de sus utilidades sobre la base observable de sus decisiones o, lo que es lo mismo, un comportamiento racional, de tal forma que este teorema de representación convierte la teoría de la decisión en una teoría empírica predictiva[3]. La teoría de la decisión se presenta en un formato axiomático (Luce y Raiffa, 1957, entre otros muchos) que recoge los criterios formales de racionalidad y permite examinar con rigor las críticas, aunque al precio de alejar la noción de racionalidad de la intuición

3. Una de las fuentes principales de crítica a la teoría de la elección racional proviene de la constatación de que en algunos contextos la función puede tener varias soluciones racionales, e incluso puede que no exista ninguna solución racional (Elster, 1989). Estos descubrimientos limitativos pueden entenderse alternativamente como limitaciones intrínsecas de la racionalidad o como limitaciones predictivas de nuestra teoría del comportamiento racional, dependiendo en qué aspecto de la noción de racionalidad hagamos hincapié, el normativo en el primer caso, el descriptivo en el segundo.

común no preparada. Lo mismo ocurre con las normas lógicas que regulan las inferencias válidas desde la perspectiva de la racionalidad epistémica: la lógica se ha convertido ya en una rama muy abstracta de las matemáticas que nos permite descubrir y postular nuevas reglas y criticar otras siguiendo la construcción autónoma que tiene en tanto que teoría formal. Sólo en último extremo nos preocupamos de saber si nuestra teoría formal es coherente con los hechos, a saber, con los razonamientos que realizan los sujetos empíricos en la práctica. Resultaría, pues, que la noción de racionalidad es como un obstáculo natural en nuestro camino que tendremos que atravesar perforando túneles que parten desde diversas direcciones.

A nosotros nos interesan las aportaciones a esta tarea que provienen tanto de la observación empírica de los sujetos cuando razonan, como de otras consideraciones filosóficas más generales. Desde ambos puntos de vista, el modelo es limitado pero no incorrecto. Los filósofos que se ocupan de la teoría de la acción han señalado varias limitaciones y dificultades, aunque no todas sean relevantes para analizar la racionalidad. Consideraremos en lo que sigue tres clases de dificultades que una teoría de la racionalidad humana debe superar:

1) La evidencia empírica de que los sujetos no parecen adecuarse siempre a las normas de racionalidad que predice el modelo, y, por el contrario, muestran un conjunto de sesgos y defectos sistemáticos que han sido bien corroborados por los psicólogos. El modelo, a la vista de estos datos, aparece como una idealización que no tiene en cuenta las circunstancias o condiciones normales en las que los sujetos realizan sus inferencias, por lo que debe ser complementado con una teoría funcional de esas circunstancias.

2) El hecho de que, teniendo en cuenta el nivel personal y no simplemente el intelectual, el modelo parece excluir de la racionalidad todo tipo de componentes emotivos, tan importantes como son en la producción y control de la conducta.

3) El carácter interpersonal de la racionalidad: puesto que la racionalidad es una virtud relacional, no depende solamente de las circunstancias que ofrece la situación de inferencia, sino también del realizarse en una comunidad de sujetos racionales que interaccionan con el agente.

4) Por último, nos encontramos con las preguntas que más discusiones filosóficas han suscitado: ¿cómo justificar las normas de racionalidad?, ¿cómo fundamentar racionalmente la racionalidad?, ¿es la autofundamentación un límite insalvable para la racionalidad?

En las siguientes secciones trataremos sucesivamente estas dificultades.

III. DE LAS PERVERSIONES DE LA RACIONALIDAD Y EL PROBLEMA DE SU EXPLICACIÓN

La gente no siempre se comporta racionalmente, como todo el mundo sabe. En lo que no se suele reparar es en que el comportamiento irracional plantea un serio problema, puesto que la atribución de racionalidad está ligada como criterio suficiente a la descripción intencional de la conducta: no es posible atribuir intencionalidad al sujeto si previamente no le suponemos racional. De modo que los fallos graves o sistemáticos de la racionalidad estarían poniendo en cuestión la propia conducta intencional. De los niños o de los autistas no decimos que se comporten irracionalmente, sino que ponemos en cuestión la estructura intencional de su mente. Una vía para evitar el problema es la separación tajante de la intencionalidad y la racionalidad, tal como hacen algunos conocidos ensayos de nuestro entorno, como son Mosterín (1978) y Quintanilla (1981), en los que, por definición, se postula como condición de racional el ser previamente intencional. Así, Quintanilla (1981), siguiendo simultáneamente el emergentismo de M. Bunge y el origen piagetiano de las operaciones lógicas, acepta la siguiente definición: «Un animal posee inteligencia lógica si es capaz de pensar y algunos de sus pensamientos tienen la estructura de un retículo distributivo y complementado» (149). La racionalidad sería, pues, un orden tardío introducido sobre las representaciones, y el comportamiento irracional sería fruto de una decisión de apartarse de este orden de pensamientos: «Nuestros sistemas neuronales funcionan de tal manera que a veces inventamos nuevas teorías y, por lo general, estamos siempre aprendiendo de la experiencia. Desde luego siempre existe la posibilidad de que uno se niegue a pensar racionalmente o a aceptar la crítica racional. Pero es más plausible pensar que nuestras mentes seguirán trabajando a favor de la razón». Mosterín (1978), en la misma línea, explica claramente cuál es la naturaleza de la racionalidad que nos permite hacer esta división: «La racionalidad creencial no es una facultad más o menos misteriosa que unos tendrían y otros no. La racionalidad creencial es un método a la disposición de todos [...] La racionalidad práctica, al igual que la teórica, no es una facultad psicológica, sino un método, una estrategia» (51-53). Esta concepción regularista, que sitúa la racionalidad en el mismo plano que otras reglas morales, tiene varias consecuencias paradójicas. La primera ya la expuso Platón en el *Menón*: Si la racionalidad es lo que hace que podamos aprender, ¿cómo puede adquirirse la racionalidad si no se posee ya? Otra, mucho más curiosa, es la de cómo es posible que alguien decida comportarse irracionalmente. Mientras que es fácil explicar por qué uno cae en tentaciones contra sus principios morales, no es tan fácil explicar por qué uno cae en tentaciones contra la racionalidad: si es una decisión, debe haber sido sope-

sada para ser intencional; si ha sido sopesada, lo ha sido sobre la base de razones... el regreso al infinito ya está creado. Pero aun si fuera correcto este anclaje de la razón en la voluntad, todavía nos quedan por explicar los fallos de la racionalidad, que no tienen un carácter ocasional, sino sistemático e involuntario, cuales son los que nos ocuparán en esta sección.

Diferente a esta tradición moralista es la tradición aristotélica, mayoritariamente seguida por los filósofos, que considera la racionalidad como una facultad constitutiva de la especie. Esta concepción evita dos de las graves dificultades de la anterior; sin embargo, tiene los mismos problemas para explicar la constatación empírica de nuestro comportamiento irracional. La solución tradicional, desde Aristóteles, para explicar la irracionalidad, desarrollada hipertróficamente en la Edad Moderna, ha sido la distinción entre la capacidad ideal de la razón como facultad y la realidad de su ejercicio en medio de pasiones y otras distorsiones. Mientras se postula que todos somos racionales, se admite que la interferencia de fuerzas externas, como las pasiones, prejuicios o ideologías, pueden llevar a un ejercicio erróneo de la facultad. Una versión actual de esta tradición postula la racionalidad, en analogía con la explicación chomskiana de la facultad lingüística, como una competencia que el sujeto posee como fruto de su desarrollo neuronal y que se muestra externamente en su actuación inferencial (Cohen, 1981). En el mismo espíritu del racionalismo tradicional, esta concepción explica los errores de irracionalidad por la interferencia de agentes externos. Si estas dos concepciones son adecuadas para dar cuenta de la irracionalidad, es algo que debemos sopesar a la vista del catálogo de nuestros fallos en la racionalidad.

1. *Los errores del pensamiento cálido*

Son los errores más conocidos y tratados por los filósofos ya desde la época platónica. Los llamamos errores del «pensamiento cálido» para señalar que están involucrados en ellos efectos motivacionales, a diferencia de los otros que son errores meramente intelectuales. Los más importantes son:

— *Akrasia*, incontinencia o debilidad de la voluntad: consiste en realizar una acción contra el mejor o más racional juicio propio. El sujeto desea un objetivo, cree que es posible realizarlo mediante un acción, y, sin embargo, no la realiza. El fumador convencido de que debe dejar de fumar, el estudiante que es incapaz de levantarse para ir a clase, ejemplifican la *akrasia*, tan familiar, por otra parte, a todos.

— *Autoengaño* o racionalización de la propia conducta. Muy estrechamente relacionada con la *akrasia*, esta perversión de la razón es especialmente importante en la historia por haber sido el mecanismo po-

pularizado por las «filosofías de la sospecha»: el marxismo, el psicoanálisis y la moral como voluntad de poder postulan que el mundo subjetivo de razones de los individuos es mera apariencia a la que subyacen otras verdaderas razones que el sujeto no se autoconfiesa, pero que la mente limpia de autoengaño del filósofo sí puede descubrir (en las otras mentes).

Desde Sócrates se han discutido estos defectos y desde entonces encontramos tres actitudes hacia ellos: *a*) que no existen tales debilidades, puesto que es metafísicamente imposible actuar contra el propio juicio o autoengañarse conscientemente. Es la solución griega (Sócrates, *Protágoras* 358d; Platón, *República* 439e-440b; Aristóteles, *Etica Nicomaquea*, 1152a 25-27; actualmente, entre otros, Johnston, 1988); *b*) que la mente es un complejo de subsistemas que en ocasiones actúan separadamente y producen la incapacidad de acción (Davidson, 1981, Elster, 1983); *c*) que se trata de un auténtico defecto irracional y de mal funcionamiento del sistema deseos-creencias, pero que tienen un grado de motivación en los factores y procesos que se necesitan para explicar la conducta humana, más allá de los deseos y creencias, como son los propiamente motivacionales, que disponen de su propia autonomía (Mele, 1987; Pears, 1984).

— *Efectos de la interacción causal aberrante entre deseos y creencias.* Han sido popularizados por Elster (1983, 1984), a causa de su importancia en la acción social. Son mecanismos que violan el requisito de la mutua independencia de los deseos y las creencias. En ellos una de las instancias actúa sobre la otra cambiando el contenido o la fuerza causal. Los dos más importantes son el *pensamiento desiderativo* (*wishful thinking*) y el *decaimiento de la voluntad.* En el primero la fuerza del deseo de que ocurra algo tiende a sesgar el buen juicio sobre las probabilidades de que ocurra; en el segundo, el caso de la fábula de La Fontaine del zorro y las uvas, la dificultad constatada de alcanzar un objetivo actúa sobre el deseo haciendo decaer su fuerza. Explicarían, según Elster, las dificultades para desarrollar una acción colectiva racional.

2. *Los sesgos en la atribución de probabilidades en contextos de incertidumbre.*

Han sido mucho menos discutidos por los filósofos que los anteriores, ya que son menos aparatosos en apariencia como mecanismos de irracionalidad, pero, desgraciadamente, son mucho más profundos y extendidos. Se trata de mecanismos sistemáticos, que operan en los sujetos independientemente de factores externos, emotivos, sociales, culturales, etc. Son, seguramente, constitutivos, en el mismo sentido en el que lo son las ilusiones perceptivas y, al igual que ellas, tienen cierta inercia o continuidad de acción después de haber sido descubiertos. Su importancia

radica en que son mecanismos que sesgan sistemáticamente el peso que se concede a la evidencia y violan principios implícitos elementales de la probabilidad, en cuyo uso el sujeto racional se presumía diestro, incluidos aquellos contextos de acción paradigmáticamente racional, como es el caso de la investigación. Tal vez constituyan el mayor descubrimiento psicológico de nuestro siglo, pues nos muestran el camino por donde se puede a penetrar en los sistemas reales de razonamiento. Kahneman, Slovic y Tversky (1982) y Nisbett y Ross (1980) son dos, ya clásicas, recopilaciones de tales descubrimientos. Las manifestaciones más importantes de estos sesgos son las siguientes:

— *Indiferencia a la tasa base o probabilidad previa* cuando se evalúa la probabilidad de ocurrencia de un suceso. Si se comunica a los sujetos una cierta composición de una muestra, pero luego se les facilita otra información cualitativa, los sujetos tienden a pesar las probabilidades de ocurrencia de un suceso guiándose por tópicos y prototipos, que pueden haber sido activados por esa nueva información, más que por la probabilidad objetiva.

— *Insensibilidad frente al tamaño de la muestra.* Los sujetos no son muy sensibles a la representatividad objetiva de una muestra respecto de una clase y tienden a conceder representatividad general a muy pocos ejemplares observados respecto al conjunto de la clase, a pesar de que su representatividad sea muy baja. A esta familia de defectos pertenece la llamada «falacia del apostador» de la que viven todos los casinos del mundo: si sale cara tres veces seguidas, los sujetos tienden a sesgar la probabilidad de cara o cruz que en la próxima tirada sea 0.5, sin reparar en que la estadística solamente tiene significado en grandes números.

— *Efectos de prominencia en la atribución causal* y en la clasificación: los sujetos tienden a usar los rasgos más prominentes que presenta una situación observada a la hora de atribuir causas o de clasificar los hechos, en vez de comprobar todas las propiedades o sucesos que ocurren en la situación para analizar los flujos causales o atribuir correctamente los papeles. Es el efecto que, desgraciadamente, nos hace, en muchas ocasiones, ver a las víctimas como culpables.

La formulación de estos defectos parece un poco esotérica, pero sus efectos son desastrosos. La publicidad vive de ellos por el carácter sistemático e involuntario de las inferencias dirigidas por estos sesgos. Pero también la propaganda política, la atribución de causas y efectos sociales y, lo que es más grave, se han observado también en el propio sistema científico. Varios de los efectos kuhnianos de resistencia al cambio de paradigma y de generación de autoridades internas pueden estar sesgados por algunos de estos defectos.

3. *Sesgos en el razonamiento lógico: el problema de las cuatro tarjetas.*

Wason y Johnson-Laird descubrieron a comienzos de los años setenta que, cuando se enfrentaba a los sujetos a un caso simple de *modus tollens*, aunque presentado en términos abstractos, los sujetos eran incapaces de aplicar la regla correcta en un altísimo y sistemático porcentaje. En el famoso experimento se presentan cuatro tarjetas con una letra o un número en cada cara observada con la siguiente secuencia A, D, 4, 7. Dado que por la otra cara pueden tener escrito otra letra o número, se les pide que comprueben la validez de la regla, «cuando en la primera cara hay una vocal, en la cara opuesta hay un número par». Para ello deben levantar el número mínimo de tarjetas suficiente para comprobar la regla. Entre el 60% y el 80% de los sujetos no avisados eligen A y 4, en vez de la solución correcta A y 7. Al igual que en los casos anteriores, también hay un efecto de persistencia del defecto: los sujetos presentan una gran dificultad para entender el error, cuando se les explica en términos abstractos. A pesar de que cuando se presenta la tarea con situaciones y regularidades familiares para el sujeto la tasa de errores disminuye, no hay ningún consuelo, dado que se supone que la regla del *modus tollens* es la estructura básica del pensamiento crítico, que, de acuerdo a la doctrina oficial de la racionalidad, debería ser independiente del dominio al que se aplica. De todos los resultados probablemente sea el más catastrófico, dado que afecta a los mecanismos más básicos de inferencia natural.

El estudio de los errores y sesgos en el razonamiento se ha convertido en el más importante de nuestros instrumentos para el estudio de la racionalidad. Desde el punto de vista de los psicólogos, los sesgos son una ventana abierta a las estructuras primigenias de nuestra mente mucho más relevante que la de los razonamientos válidos, puesto que nos informan directamente sobre las restricciones reales bajo las que se desarrolla el pensamiento, sobre los mecanismos de acceso a la memoria, sobre la cantidad de información tenida en cuenta en la memoria a corto plazo, sobre el tiempo de computación y sobre el sistema de evaluación de la propia inferencia (Johnson-Laird, 1983, 1990; Cherniak, 1986). El estudio de las restricciones es también el medio para desarrollar nuevos modelos formales de razonamiento que recojan o simulen el razonamiento natural. Un modo de interpretar estos errores, y sobre todo el hecho de que se produzcan sistemáticamente, es considerar seriamente la analogía con las ilusiones perceptuales (Gomila, 1993). Desde este punto de vista, la racionalidad se explicaría evolutivamente como una función que se desarrolló primitivamente en ciertas condiciones normales de tratamiento natural de la información. De esta manera se explicaría que, cuando el problema exige mayor abstracción, o cierto

cuidado matemático que no es el del razonamiento común, se produzcan errores y distorsiones similares a las que se producen en nuestros sentidos cuando, por ejemplo, modificamos artificialmente los tamaños relativos de las cosas o sus distancias, de manera que sesgamos nuestro cálculo natural de perspectivas. Esta explicación es correcta y, sin embargo, no deja de parecer incompleta, dado que existe una profunda disanalogía con la percepción visual: aunque hay un efecto de inercia en los sesgos, de manera que la información falsa persevera en su acción después de comunicarle al sujeto los resultados, a diferencia de la percepción visual, el sujeto avisado no comete errores en la siguiente ocasión. El problema de las cuatro tarjetas funciona estadísticamente, pero solamente lo hace una vez. Dicho de otra forma, la racionalidad cumple también una función de segundo orden que es la de control de calidad de las propias inferencias (Broncano, 1993). Lo intrigante de una concepción funcional simple es esta capacidad de automodificación y aprendizaje de las normas de racionalidad. A diferencia de otras funciones biológicas, como son las perceptivas, el cerebro presenta una doble plasticidad, de especie, a través de la evolución, y de individuo, a través del aprendizaje cultural. Como especie disponemos de una racionalidad mínima (Cherniak, 1986; Johnson-Laird, 1983; Stich, 1990) que nos permite asignar contenidos intencionales a los otros y predecir su conducta. Como especie también somos capaces de construir reglas con nuevas funciones naturales de control sobre los procesos informacionales. Para asignar un papel a las ilusiones cognitivas necesitamos al tiempo una teoría extendida de las funciones mentales, que incorpore la racionalidad y una teoría extendida de la decisión racional (Nozick, 1993) que incorpore, junto los elementos heredados de nuestro diseño cognitivo, los elementos aprendidos en el desarrollo cultural.

IV. LA RACIONALIDAD DE LAS EMOCIONES

Las emociones deberían figurar necesariamente en cualquier esquema funcional de la racionalidad humana, ya que constituyen un sistema motivacional importante, paralelo al sistema de creencias y deseos, con el que interactúa en la producción de la conducta[4]. Los filósofos modernos

4. El sistema afectivo está constituido esencialmente por la amígdala, un órgano del cerebro medio con proyecciones desde el núcleo talámico. El sistema recibe información multimodal sensoria y la proyecta a las áreas de reconocimiento como el lóbulo frontal. El procesamiento cognitivo es independiente de este sistema. El hipocampo es en este caso el encargado de la integración multimodal y el envío a las zonas de reconocimiento. Hay, sin embargo, varias formas de interacción entre ambos subsistemas. El hipocampo modula la información relevante para el sistema afectivo y, después del reconocimiento, hay una realimentación al sistema afectivo, probablemente con el objeto de perfilar conscientemente la emoción (Le Doux, 1989).

llamaron a las emociones pasiones del alma, para subrayar el carácter pasivo del sujeto respecto a las pasiones [5], quizás porque estaban excesivamente sensibilizados hacia la falta de control que parecen mostrar algunos procesos emocionales. La psicología contemporánea, sin embargo, ha subrayado los aspectos motivacionales que existen en las emociones y, por consiguiente, su función dentro del sistema de producción de la conducta e incluso el desarrollo intelectual [6]. Es en este contexto funcional en el que las consideraciones de racionalidad de las emociones han comenzado a recibir la atención creciente de la filosofía. Varios filósofos han subrayado el carácter intencional de las emociones (De Sousa, 1986; Solomon, 1980; Gordon, 1987; Lyons, 1980), puesto que puede decirse de muchas de ellas que tienen objeto intencional o contenido (miedo a las arañas), lo que hace su descripción estructuralmente similar a la de las actitudes proposionales [7]. Quizá sea excesivo afirmar que tienen estructura cuasi-proposicional (Johnson-Laird y Oatley, 1992), pero lo cierto es que son una parte esencial de la conducta intencional y la estructuración del comportamiento social, por lo que tienen una función no menor que la que realiza el lenguaje.

En la historia de la filosofía encontramos tres modos de tratamiento de las emociones. En primer lugar, una aproximación fenomenológica, puesto que, análogamente a las sensaciones, los aspectos cualitativos parecen esenciales en la dinámica de las emociones. Es el estilo dominante entre los filósofos y, sobre todo, el que nos muestra frecuentemente la literatura. En segundo lugar, una aproximación conductista-fisiológica, que correlaciona sucesos que causan emociones con cambios fisiológicos o conductuales. Esta es, por ejemplo, la aproximación inaugurada por Darwin, quien estudió la similitud de los gestos emotivos entre los animales y el hombre. Existe, por último, una aproximación funcional, en la que se comienza por conjeturar cuál puede ser la función evolutiva del sistema emocional para después establecer las categorías en las que debe situarse. Cada una de estas aproximaciones producirá probablemente su propia clasificación y una división propia entre emociones básicas y emociones derivadas. Por otra parte, encontramos una actitud hiperracionalista que equivale simétricamente a la actitud romántica irracionalista, que huye de cualquier intento de establecer puentes entre las emociones y la racionalidad. La perspectiva funcionalista, que no es

5. Lyons (1980) es uno de los escasos, y sin embargo accesible, estudios históricos sobre las concepciones filosóficas de la emoción.

6. De hecho ésta es la razón de que llamemos a este conjunto de sucesos emociones, término cuyo origen latino es *emovere*, que tiene connotaciones de causa de movimiento. La observación de la autofenomenología de las emciones nos enseña mucho sobre la falta de transparencia de la mente.

7. En una oración de actitud proposicional como «Alicia cree que El Señor de los Anillos es Gandalf» distinguimos la actitud propiamente dicha como *cree* que, de su contenido, «el Señor de ...». Este mismo análisis sería el aplicable a las emociones, distinguiento el modo de la emoción de su objeto.

ajena ni al tratamiento fenomenológico ni a la descripción fisiológica (externa o interna), tampoco desprecia la idea de que muchos elementos de las emociones sean ajenos al control racional de la conducta; sin embargo, considera prioritario el análisis de cuáles pueden ser las funciones asumidas por el sistema emocional dentro del conjunto de nuestros sistemas de control.

Las emociones conectan con la racionalidad en la medida en que forman parte, de alguna manera, del sistema cognitivo y no son meros mecanismos reflejos. Dentro de nuestros sistemas cognitivos podemos distinguir aquéllos cuya función se realiza en el flujo de la información, en su almacenamiento o recuperación, y aquéllos cuya función es el control del flujo de la información, con objeto de garantizar que se cumplan los fines del organismo. Si las emociones se relacionan con el sistema cognitivo es precisamente a través de esta segunda función. Por otra parte debemos tener en cuenta que la función evolutiva, la función para la cual fue seleccionada una cierta característica, no tiene que coincidir necesariamente con la función o funciones (o con la totalidad de ellas) que realiza actualmente en el organismo. Posiblemente la distinción tradicional entre emociones básicas y emociones derivadas tenga que ver con la diferencia de funciones asignadas en la evolución biológica y funciones adquiridas en la evolución cultural.

Se pueden distinguir estructuralmente tres diferentes aspectos en las emociones. En primer lugar, el propiamente emotivo o fenomenológico. Las emociones tienen aspectos cualitativos o modales, que probablemente están relacionados con la química del cerebro, especialmente con ciertas hormonas, como la testosterona y con ciertos neurotransmisores como las dopaminas y endorfinas. Esta dependencia explica que las emociones estén conectadas con el funcionamiento del sistema endocrino, por una parte, y con ciertos subsistemas de nuestro cerebro medio como son la amígdala y el núcleo talámico. La dependencia del contenido fenomenológico de estas sustancias induce a pensar que existe una correlación nomológica entre la presencia de alguna de estas sustancias y la experiencia emocional, aunque el sujeto en ocasiones distingue entre emociones que están ligadas al mismo aspecto neurofisiológico[8]. En segundo lugar está el aspecto cognitivo de las emociones. Las áreas prefrontales del cerebro están conectadas con las funciones de valoración de los aspectos prominentes en un suceso para los fines del sujetos. El sistema emotivo, por su parte, parece estar encargado de la función de hacer que algunos objetos, propiedades o sucesos sean especialmente resaltados. Las emociones sirven de filtros detectores de intereses y peli-

8. En unos famosos experimentos, en los que se pedía a los sujetos que midieran su grado de excitación sexual ante imágenes eróticas, estando en reposo y después de un ejercicio fuerte, se observaban diferencias muy sustanciales que eran atribuibles a la confusión del sujeto sobre su propio estado emocional, por la sola diferencia de estar o no realizando ejercicio físico.

gros (externos e internos), así como de la consecución de objetivos intermedios. En tercer lugar está el aspecto ligado al control de la conducta. Aunque las emociones tienen una dimensión interna resaltada por los aspectos anteriores, su principal función biológica está ligada al control de la conducta. Así, una emoción como el miedo puede manifestarse de diferentes formas en el desarrollo de un primate en una fase, como conducta de llamada; en una siguiente, como conducta de inmovilidad; y en una tercera fase, como conducta de lucha o de huida. La manifestación en una de esas diferentes conductas depende de cómo la emoción del miedo interactúa con los demás sistemas cognitivos, cumpliendo, pues, diferentes papeles funcionales[9].

Las emociones son sistemas de control de la información procesada por un sistema cognitivo que debe atender a intereses múltiples y complejos en un medio ambiente incierto (Johnson-Laird y Oatley, 1992), donde la obtención de información es un proceso lento y costoso. Nuestro sistema deberá tener ciertas disposiciones o capacidades para seleccionar o sintonizar los estados externos o internos relevantes para esos intereses. La función de este sistema de detección es indicar que se tiene que hacer algo y qué es lo que habría que hacer. Con este objetivo se activa un pequeño repertorio de esquemas de acción posibles, como es el detener la acción, huir o, por el contrario, continuar el plan, etc. En los animales sociales, las emociones cumplen la función añadida de servir de ajuste entre la conducta cooperativa o competitiva de los otros, puesto que la cooperación hace necesario coordinar planes y acciones situadas (Broncano, 1990).

Respecto a la racionalidad, son sistemas que cubren las lagunas de la racionalidad imperfecta al servir como filtros de información de un sistema de recursos limitados. Parten el mundo en categorías muy simples de sucesos y activan acciones que cubren muchos intereses. Forman parte, pues, del mismo sistema que la racionalidad. Eso no las convierte en necesariamente racionales. Las emociones interfieren a veces bajo condiciones de estrés el desarrollo del hipocampo y producen marcas permanentes en el aprendizaje, como la angustia, el miedo o las fobias, que son claramente irracionales y disfuncionales. También ocurre lo contrario, ya que el sistema de realimentación del cerebro al sistema emotivo maduro es un sistema esencial de reforzamiento del aprendizaje. Por último, hay un sentido derivado en el que se puede hablar de racionalidad de las emociones (De Sousa, 1987) y es el que tiene lugar en la medida en que el desarrollo cognitivo permite una realimentación y maduración del sistema afectivo, que se expande y convierte en un filtro más

9. Kalin (1993) es un magnífico y sugerente estudio, en la perspectiva antropológica, de la conducta de miedo en las crías de chimpancé, nuestro más cercano pariente evolutivo. El hecho de que la conducta de miedo se encuentre estereotipada en distintas fases de desarrollo del bebé dice mucho sobre la funcionalidad biológica de esta emoción.

fino y mejor sintonizado a los objetivos más desarrollados y maduros del individuo. Una mala integración de ambos sistemas es la fuente más segura de irracionalidad cognitiva y conductual.

Hay una división entre emociones básicas y derivadas, proveniente de funciones básicas que, seguramente, cumplen las emociones. Johnson-Laird y Oatley (1992) proponen las siguientes: la *felicidad*, ligada a la recompensa interna por haber alcanzado fines o progresar en ellos; la *tristeza*, que cumple el objetivo contrario; la *angustia*, ligada al bloqueo de planes; el *miedo*, producido por el conflicto entre objetivos o por una situación de amenaza, y el *disgusto*, ligado a la percepción de algo rechazable. Es una cuestión empírica la clasificación y el estudio de estas emociones. También lo es su carácter o no transcultural. Así, aunque los términos de emoción entre distintos idiomas sean intraducibles en contextos de traducción radical, si es cierta la hipótesis del carácter básico, la empatía de especie pudiera resultar un mecanismo más profundo que la traducción y asignación de creencias en contextos radicales.

Repárese en que esta concepción funcionalista de las emociones es compatible con la idea de que sean simulables. Mientras muchos de sus aspectos cualitativos son producto de la química especial del cerebro (que es simulable por otros medios, mediante la sustitución de neurotransmisores y hormonas por otras sustancias como las drogas), los aspectos funcionales son independientes de su instanciación en un sistema orgánico como el nuestro. Un sistema complejo, de intereses complejos y que deba tomar decisiones usando una memoria de trabajo con recursos limitados, tendría que desarrollar necesariamente un sistema similar al emocional, de manera que el sistema emocional no es necesariamente incompatible con una concepción funcionalista (biológica) de la mente, pero no hay nada que impida que las funciones biológicas sean simulables.

V. LA FUNCIÓN SOCIAL DE LA RACIONALIDAD Y LA HIPÓTESIS DEL ANIMAL MAQUIAVÉLICO

La importancia que hemos concedido a la adscripción de racionalidad a los otros, no sólo como instrumento predictivo de la conducta ajena, sino también como criterio suficiente de racionalidad, recibe un sólido fundamento de las condiciones sociales evolutivas y epigenéticas en las que se formó la inteligencia, lo que arrojaría una nueva luz sobre la función de la racionalidad. La racionalidad es una capacidad innata que desarrollan por epigénesis los cerebros normales humanos. La interacción social con otros miembros del grupo es necesaria, pero no suficiente para el desarrollo de esta capacidad. La estructura social de nuestra especie es, sin embargo, algo más que una condición de medio ambiente favorable

en la evolución de la racionalidad. Varios primatólogos han sostenido recientemente la tesis de que la inteligencia social pudo ser un factor esencial en la separación evolutiva de nuestra especie[10]. Sabemos que los primates desarrollan una fuerte estructura social: la formación de grupos con cierta estructura social daría una ventaja evolutiva para los individuos que pertenecieran al grupo, puesto que incrementaría notablemente las posibilidades de aprendizaje individual por imitación, ofreciendo una oportunidad para el desarrollo de habilidades técnicas que se transmiten por observación (Goodall, 1989; Mosterín, 1993), por consiguiente, para el desarrollo de una auténtica cultura que ya no se transmite genéticamente, sino por enseñanza o aprendizaje. Es sorprendente, pero no casual, que, incluso entre los primates, las tecnologías básicas de subsistencia y la formación de grupos están estrechamente relacionadas. Sin embargo esta misma ventaja se convierte en un problema cuando queremos explicar la función biológica de la inteligencia, suponiendo que por ésta entendemos la capacidad para resolver creativamente los problemas que presenta el entorno o la propia mente.

El proceso de formación de la razón ha sido un proceso relativamente rápido, hablando en términos evolutivos. No hay mucha coherencia entre esta rapidez y el equilibrio de conductas que produciría la existencia de bandas con estructura social y tecnologías básicas de subsistencia, dado que el cambio del medio ambiente no ha sido tan importante como para explicar la génesis de la racionalidad. ¿Qué premio pudo haber para que un individuo desarrollase soluciones creativas dentro de una estructura de grupo? Como ocurre en los sistemas burocráticos, la creatividad y la socialización parecen estar reñidas. A menos que consideremos el proceso de socialización, precisamente, como la fuente de problemas que conduce al desarrollo de la inteligencia. Ocurre, sin embargo, que sí existe una fuerte presión, y una recompensa ligada a ella, para modificar el propio *status* dentro del grupo[11]. Los machos deben resolver el problema de emparejarse con las hembras en el período de estro, las hembras deben obtener comida para ellas y sus crías, que se reparte proporcionada y ordenadamente según el puesto jerárquico; todos deben resolver a su favor la agresividad de los machos dominantes o, en su caso, de los que aspiran a serlo. La hipótesis de una inteligencia social cobra ahora un nuevo sentido: la creatividad consiste esencialmente en manipular la conducta de los otros para los fines propios .

Hay dos maneras de conseguir esta instrumentalización. Una es la mera manipulación de la conducta de los otros. La segunda es la manipulación de sus mentes, mediante la representación interna propia de cuál

10. Ver Humphrey (1989), Jolly (1989), Cheney y Seifarth (1990).

11. Chance, y Mead (1988) establecen medidas cuantitativas del tiempo empleado por bandas de primates en las diversas actividades relacionadas con el comportamiento de supervivencia y sociabilidad.

es la representación interna de las creencias y deseos del otro, y el desarrollo de un plan de acción consecuente. Si un individuo es capaz de predecir la conducta del otro en virtud de sus representaciones internas, también puede manipularla, manipulando sus representaciones; por ejemplo, suministrándole información incorrecta o favorable a los propios fines (Premack y Woodruf, 1978; Dennett, 1983; Perner 1989). Mientras que la manipulación de conductas puede desarrollarse simplemente por alguna forma de condicionamiento complejo, como ocurre con la domesticación de animales, la manipulación de las mentes exige mecanismos de reconocimiento mutuo. Estos sistemas no están basados simplemente en mecanismos afectivos o emocionales, sino en la atribución mutua de estados mentales y en la capacidad de realizar inferencias basadas en estos estados. Fue el desarrollo progresivo de esta capacidad el motor de la inteligencia, en cuanto se unió a la transformación de las formas y estructuras sociales. Premack y Woodrof (1978) llamaron «teoría de la mente» a esta capacidad de atribuir al otro estados mentales y predecir la conducta. Dennett (1983) indicó que existen grados de atribución que están determinados por el orden de profundidad de la atribución. En el primer orden atribuimos un contenido como X quiere O, en el segundo orden, una atribución encajada como «X cree que Y quiere O», etc. La inteligencia social y la génesis de intenciones comunicativas exige probablemente atribuciones de tercer o más orden (Gómez, 1989). No sabemos hasta qué grado han llegado primates superiores como los chimpancés. Perner (1989) observó que los niños adquieren esta capacidad entre los tres y cuatro años y Leslie, Frith y Baron-Cohen (1985) han detectado un defecto en esta capacidad entre los autistas, defecto, de origen posiblemente genético, que explica el extraño síndrome del autismo. Los niños autistas no son capaces de comprender todas aquellas conductas de los otros en los que están implicados sus estados mentales, en forma de deseos o creencias, de manera que quedan afectadas partes fundamentales de la conducta social que exigen esta capacidad. Harris (1992), por otra parte, ha estudiado cómo el niño va formando progresivamente la distinción entre deseos y creencias, quizás a partir de estados intermedios de atribución emocional. El desarrollo de la atribución distinta de deseos y creencias es, a su vez, esencial para el desarrollo de la racionalidad como sistema de control de inferencias: los planes pueden desarrollarse y hacerse complejos solamente cuando pueden distinguirse metacreencias y elaborar juicios de compatibilidad entre objetivos y prioridad de actuaciones.

Aunque la comprobación de esta hipótesis es una cuestión empírica, la estructura funcional de la racionalidad, ligada al control y predicción de planes complejos, originados por la acción colectiva, es algo más que una hipótesis empírica; es, probablemente, el mejor análisis que tenemos de la naturaleza de la racionalidad. Pero aún nos queda la pre-

gunta más ardua: ¿Por qué ser racionales?, ¿cómo fundamentar los aspectos normativos de la racionalidad?

VI. LA FIABILIDAD DE LA RAZÓN Y EL EQUILIBRIO REFLEXIVO AMPLIO

Son muchos los filósofos que han desesperado de la posibilidad de fundamentar la racionalidad: los escépticos en la tradición humeniana, como Stich (1990) y los racionalistas críticos como Popper. Todos ellos creen que cualquier intento de fundamento está viciado de circularidad, o el fundamento, por ser más débil que lo que trata de fundamentar, nos lleva a un regreso al infinito de fundamentaciones. Otros filósofos, por la magia de los argumentos transcendentales, opinan que no podemos no ser racionales, pero no alcanzan a explicar por qué no lo somos en tantas ocasiones. El que no sea posible fundamentar las normas de racionalidad es el gran escándalo de la filosofía contemporánea, como en otros tiempos lo fue la imposibilidad de demostrar la existencia del mundo externo. Hace dos décadas se abrió una puerta a la esperanza cuando J. Rawls, en su *Teoría de la justicia*, generalizó y divulgó el método del equilibrio reflexivo de Nelson Goodman para justificar normas o reglas de inferencia. Según este método, aceptamos una norma de inferencia porque produce inferencias que intuitivamente consideramos válidas o aceptables y, de otro lado, consideramos válidas las inferencias que sean el producto de una regla que ya hemos aceptado. El método es circular, pero no viciosamente circular: el equilibrio reflexivo resulta de la interacción temporal entre los dos polos del equilibrio, la intuición y las normas. Podemos suponer que, a medio plazo, las normas que tiendan a producir juicios o inferencias inaceptables para la intuición, serán rechazadas, y, por otra parte, las inferencias realizadas por el sujeto tenderán a adecuarse al patrón establecido por las normas, de donde podemos deducir que un equilibrio, aunque sea inestable, es también una justificación de nuestras normas.

A pesar de que esta estrategia resulta atractiva, se abren algunas cuestiones que impiden que el método del equilibrio reflexivo, tal como fue formulado por Goodman, genere una justificación suficiente de nuestras normas de racionalidad. La primera deriva del carácter individualista que caracteriza el método. El equilibrio se produce entre las intuiciones y las normas que aceptan los individuos en tanto que individuos, pero no hay ninguna garantía de que sean universalizables. Stich (1990) recoge una anécdota de Nisbett, uno de los psicólogos especializados en el estudio de los sesgos en las inferencias. Cuando Nisbett exponía sus resultados en audiencias de psicólogos o filósofos, además de las objecciones dirigidas a poner en duda que los individuos estuvieran realmente equivocados, dado su leal entender, cuenta Nisbett que siempre había alguna

objección de fondo como: ¿por qué está usted seguro de que su norma de asignación de probabilidad es la correcta?, o ¿qué es lo que hace de una inferencia una inferencia realmente buena? Puesto que si la intuición es una facultad individual, cada cual podría generar un sistema de inferencias adecuado a sus necesidades sin coincidir necesariamente con el de los demás. El problema de Nisbett es un problema de universalizabilidad de los resultados del método del equilibrio reflexivo, lo que, por otra parte, es esencial para que la norma tenga valor como norma. Podríamos, quizás, defendernos de este obstáculo arguyendo que la facultad de la intuición produce resultados universales, tal vez porque en sí misma esté dotada de universalidad, tal como sostenía la versión cartesiana del sentido común —quizá porque existan capacidades universales de especie—, pero Stich (1990 y 1991) ha señalado convincentemente cuál es el error que cometemos al tomar este sendero: la intuición será un mecanismo universal de evaluación solamente si es independiente de la información que procesa y valora, pero si, por el contrario, se diese el caso de que la propia intuición estuviese conformada por ciertos ejemplares paradigmáticos de inferencia, y por consiguiente se aplicase tomando esos casos como muestra y guía de evaluación para otros similares, de manera que las propias reglas tuviesen como sistemas de referencia esos prototipos, dejaría entonces de ser un mecanismo universal, dado que, como facultad, dependería de la historia cognitiva del sujeto y de las inferencias que dicha trayectoria ha llevado a conformar como ejemplos paradigmáticos de racionalidad. Y precisamente los psicólogos argumentan que los sujetos, al razonar intuitivamente, emplean reglas rápidas basadas en prototipos o heurísticas, lo que parece apoyar empíricamente la sospecha de Stich. De manera que no es muy confiable la idea de argumentar postulando una capacidad innata de acertar en la evaluación de inferencias, porque pudiera ser el caso de que tomásemos lo que no es más que un producto contingente de nuestra historia cognitiva. Quizás no esté fuera de lugar realizar una analogía con la estructura euclideana del espacio, que, según Kant, formaría parte de las condiciones *a priori* de la organización de toda experiencia posible, cuando seguramente no sea más que un producto de nuestra dotación evolutiva de la percepción. Como sabemos, en el siglo XIX aparecieron otras geometrías alternativas a la euclideana que eran en sí mismas consistentes, mostrándose el carácter contingente de la geometría euclideana.

Una segunda vía que podemos intentar es la de buscar una explicación evolutiva a la racionalidad. La racionalidad se justificaría desde esta perspectiva, porque el hecho de que cumpliese una cierta función de control de inferencias explica el que fuese seleccionada como un rasgo de nuestra especie. La idea de fondo es considerar que las propiedades normativas de la racionalidad son un resultado de la historia evolutiva de la especie: el que nuestros ancestros se atuvieran a razones produjo una ven-

taja biológica a los individuos que desarrollaran la capacidad de controlar y mejorar las inferencias, lo que explica el que nuestro cerebro esté conformado de tal forma. De manera que la normatividad deriva del éxito evolutivo que pudieron inducir las normas de racionalidad. Sober (1981) ha explicado una forma en la que pudo operar la selección sobre un sistema de racionalidad determinado. Un sistema cognitivo, como el de nuestros ancestros evolutivos, es un sistema que aprovecha la información del medio para obtener satisfacción a sus necesidades de tal modo que el incremento en la información obtenida está relacionado con el éxito en la satisfacción de necesidades, dando por supuesto que este incremento se produce armónicamente con las constricciones estructurales bajo las que tiene que operar el sistema. Por ejemplo, debe actuar en un tiempo limitado, y en general lo tendrá que hacer muy rápidamente, dado que una pérdida innecesaria de tiempo en obtener información convierte a ésta en inútil o contraproducente. Y una de las más importantes fuentes de pérdida de tiempo se producen al recuperar la información que se almacena en la memoria. Por otra parte, nuestro sistema tiene una memoria limitada, así que debe tener un sistema de selección de lo que se almacena, o pronto se llenará de información inútil o irrelevante (Clark, 1989, cap. 4). En estas condiciones, el guiarse por reglas abstractas, un silogismo disyuntivo pongamos por caso, puede significar una ventaja evolutiva para quienes consigan obtener más información en el plazo de tiempo más pequeño.

El dejar en las manos de reglas la producción de inferencias cumple una función biológica interesante para quienes generen esta capacidad (Nozick, 1993). Ahora bien, las reglas son seleccionadas por el éxito en la acción de quienes las siguen, y el éxito conductual no depende necesariamente de que las reglas sean perfectas desde el punto de vista lógico, sino de que consigan el máximo posible en la negociación entre las capacidades cognitivas del organismo, el tiempo de respuesta y la cantidad de información obtenida. Por ejemplo, un sistema cuyas inferencias tengan un dominio de aplicación demasiado específico, de tal modo que se aproximen a representaciones de regularidades naturales, alcanzará mucha velocidad en la extracción de conclusiones, en presencia de circunstancias antecedentes apropiadas, mas pagará el precio de la falta de plasticidad para correlacionar regularidades similares. Por el contrario, un sistema de inferencias exclusivamente abstracto deberá dedicar muchos recursos computacionales al reconocimiento de la información concreta en las circunstancias presentes, por lo que, habiendo ganado capacidad de obtención de información, pagará el precio en tiempo de reacción [12]. Es muy plausible pensar que nuestro cerebro fue diseñado

12. Pensemos en la posibilidad de contar con una biblioteca de planes de acción que se activan como rutinas ante situaciones apropiadas. De hecho la automatización de algunos tipos de conducta,

evolutivamente de forma que alcanzó un cierto grado de equilibrio, si no perfecto, sí suficiente para dar una posibilidad de supervivencia al *homo sapiens sapiens* por encima de sus otros competidores. La teoría evolutiva puede explicar de esta manera el que la selección haya operado sobre los sistmas con mejores formas de racionalidad, ya que su propio éxito les concede un cierto grado de justificación. Este equilibrio evolutivo explica, por otra parte, la existencia de sesgos[13] sistemáticos en el razonamiento. Un sistema de funciones orgánicas puede haber sido seleccionado porque las funciones conceden ventajas biológicas a los organismos que las ejercen, pero esto no implica se vaya a producir necesariamente un sistema perfecto desde el punto de vista de un diseño racional *a priori*. Por el contrario, la evolución opera al tiempo sobre estratos viejos y sobre nuevas circunstancias, genéticas o ambientales, de modo oportunista y por caminos absolutamente contingentes. La perfección biológica es solamente una ilusión del observador, producida tal vez por el hecho de que es el último espécimen de una cadena de sistemas no viables que se han quedado en el camino.

La estrategia evolutiva es sugestiva para explicar algunos rasgos normativos de nuestra dotación racional natural, pero, desgraciadamente, deja abiertas demasiadas puertas a la duda. Por una parte nos encontramos ante un problema de la familia del problema de la inducción: del hecho de que el comportamiento racional haya estado justificado hasta el momento no se infiere que haya de estarlo en el futuro. Por otra parte, el que podamos corregir los sesgos, una vez que reparamos en ellos, plantea un problema a la justificación evolutiva puesto que, si se explica cómo se han producido los sesgos, no explicamos sin embargo nuestra capacidad de transcenderlos. Al menos desde un argumento evolutivo sin más consideraciones que las tenidas en cuenta hasta el momento. Pues aunque hemos partido de la convicción de que la racionalidad es el resultado de la historia cognitiva de la especie, y que sus estructuras inferenciales son el producto de una negociación entre contenido informativo y abstracción, no hemos tenido en cuenta la plasticidad cerebral y la capacidad de aprendizaje de los individuos. La cuestión relevante es si la existencia de los individuos en comunidades de otros sujetos racionales permite modificar el curso evolutivo. En la sección anterior, por otra parte, ya hemos insistido en que, además, la principal fuente de problemas para los primates superiores: ¿puede afectar la existencia comunitaria a las pro-

como, por ejemplo, conducir coches, puede ser explicada de esta forma. La negociación entre planes flexibles y planes rápidos puede ser un punto de equilibrio fundamental en la supervivencia.

13. Cosmides (1989) y Cosmides y Toobby (1991), han encontrado asombrosos resultados experimentales usando el test de las cuatro tarjetas que parecen indicar una cierta disposición innata a encontrar más rápidamente a quienes violan las reglas del contrato social entre aportación al común y beneficio obtenido. De ser correctos los resultados de Leda Cosmides, estaríamos ante un caso de mecanismo de inferencia especializado e innato.

pias capacidades racionales? y, si de hecho hubiese modificación de las capacidades racionales, si inferir bien fuese una habilidad que pudiese adquirirse por entrenamiento, como tantas otras, ¿cómo afectaría al problema de la justificación de la racionalidad?

En un primer paso introduciremos una justificación de la racionalidad que no es completamente equivalente a la explicación histórica. En un segundo paso, recobraremos la idea del equilibrio reflexivo en términos amplios y no restringidos, lo que nos permitirá solucionar los problemas planteados por Stich.

Lo que hace importante la explicación evolutiva es que nos permite insertar lo normativo en el mundo de los hechos a través de la función que cumple la racionalidad. El esquema argumental consiste en apoyar el valor normativo de la racionalidad en el hecho de que cumple bien su función o, dicho a la inversa, que el hecho de que cumpla fiablemente su función explica que la racionalidad esté ahí[14]. Ahora bien, una forma de sostener este argumento es remitiéndonos a las condiciones evolutivas en las que emergió la racionalidad, pero no es estrictamente necesaria la dependencia histórica para establecer un criterio de fiabilidad como justificación de la racionalidad. Por el contrario, podemos elevarnos a un plano más abstracto para afirmar que el hecho de que la racionalidad cumpla fiablemente su función explica su mantenimiento[15]. Este sutil cambio desde la justificación histórica a la justificación fiabilista se explica porque lo que a nosotros nos interesa es usar explicativamente la racionalidad como una capacidad, potencialidad o virtud de las personas que se ejercita en su conducta, cuando esta conducta obtiene los resultados buscados. Por otra parte es más importante el hecho de que actualmente siga cumpliéndose la función en relación a los otros sujetos que la reconocen, así como a las circunstancias en que pudieron explicar su origen evolutivo. Esta observación es importante, además porque se trata de una función que admite grados de ejercicio virtuoso. Hay aspectos de la racionalidad que son reflejos, es decir, dependen de cómo está diseñado nuestro sistema cognitivo por la historia evolutiva. Estos aspectos contienen, como ya sabemos, sesgos importantes. Pero también hay una racionalidad reflexiva, fruto del aprendizaje y la enseñanza, que permite iluminar los sesgos y corregirlos. Así la fiabilidad de la

14. Esta distinción nos remite a dos concepciones de las funciones biológicas: la histórica o etiológica y la propensional. Adams y Ben (1992) han analizado con mucha precisión esta distinción, al tiempo que han demostrado su equivalencia desde el punto de vista ontológico, es decir, desde el punto de vista de la base causal a la que pudieran sobrevenir unas y otras.

15. Este criterio se denomina fiabilista. Comenzó siendo un criterio epistemológico para justificar el conocimiento en estos términos: el hecho de que un sistema tienda fiablemente a producir un alto grado de creencias verdaderas explica que S esté justificado al creer que p porque es el resultado de este sistema. De hecho es una aplicación a la racionalidad epistémica. Pero no existe ningún impedimento para generalizarlo a la racionalidad práctica y evaluativa.

razón depende de la excelencia de su comportamiento. Como muchas funciones, admite grados de cumplimiento, que influyen determinantemente en el carácter explicativo de esas funciones. Por ejemplo, una racionalidad mínima permite explicar simplemente el sistema de reconocimiento mutuo entre las personas, pero no podría utilizarse, por ejemplo, para explicar el asombroso éxito de las teorías científicas o sistemas tecnológicos muy complejos.

Una vez que hemos establecido esta concepción abstracta de la justificación de la racionalidad podemos volver sobre nuestro tema inicial del equilibrio reflexivo. Lo interesante del proceso es que permite una interacción continua entre la fiabilidad del sistema de reglas aplicadas y la justificación de ese sistema, ya que hace posible recoger la capacidad de aprendizaje o automodificación de la racionalidad. Pero, después de toda nuestra argumentación, no podemos basarnos solamente en un equilibrio reflexivo restringido, sino en un equilibrio reflexivo que tenga en cuenta la reflexión colectiva y las mutuas correcciones de la racionalidad a través del juego social de las conductas. Y también, cómo no, en el hecho de que la racionalidad sea más o menos adecuada respecto a circunstancias que son cambiantes, de manera que sea justificable relativamente a ellas. Incluso que lo sea «universalmente» no es contradictorio con el hecho de que pueda modificarse. No es indefendible la idea de una especie de mente oculta que podría postularse como la mano oculta en la economía clásica: mentes mal diseñadas[16] interactuando colectivamente en medio de circunstancias cambiantes, pero con la plasticidad suficiente como para autocorregirse, pueden generar, aunque no necesariamente, puntos de equilibrio en el control de la información, que identificamos precisamente con la racionalidad. No generarlos significa simplemente que el sistema camina hacia una fiabilidad menor, e incluso hacia la catástrofe. Pero nadie espera que una racionalidad humana tenga que presumir algún tipo de astucia de la razón hegeliana en la historia, que, por cierto, era consecuencia de una concepción de la evolución basada en diseños perfectos, una concepción que afortunadamente barrió la revolución darwiniana. Esta concepción amplia del equilibrio reflexivo no queda al pairo de los argumentos escépticos de Stich, puesto que es compatible con el hecho de que cada uno de los individuos pueda tener sesgos de racionalidad idiosincráticos, pero es incompatible con la sospecha de que todos (¡todos!) pudiéramos estar equivocados ahora y sistemáticamente acerca de nuestras normas de racionalidad. Este argumento, ya clásico desde Davidson, se usa para defender una concepción de racionalidad perfecta, pero entendiendo el equilibrio reflexivo en un contexto am-

16. Railton (1993) ha expuesto esta idea en el Congreso de SOFIA 1993, celebrado en La Laguna. Varias conversaciones sobre el tema con Antoni Gomila me hicieron tomarla más en serio de lo que hice en un comienzo.

plio, permite distinguir entre la racionalidad mínima suficiente para comprender la conducta de otros y la racionalidad necesaria para mantener nuestro sistema cultural o sobrevivir como especie. Que nuestra especie y nuestra cultura sean intrínsecamente irracionales es una posibilidad empírica, pero tiene tantas posibilidades de ser cierta como la contraria. Sería pedir demasiado a la racionalidad la capacidad profética.

BIBLIOGRAFÍA

Adams, F. y Enç, B. (1992), «Functions and Goal Directedness»: *Philosophy of Science*, *54*, 635-54

Baron-Cohen, S., Leslie, A. y Frithu,U. (1985), «Does the Autistic Child Have a Theory of Mind?»: *Cognition*, *37-46*.

Bennett, J. (1991), «How is cognitive Ethology Possible?», en C. A. Ristau (ed), 1991.

Broncano, F. (1990), «La acción, su razón y su circunstancia», en Ballestar (ed.), *Conocimiento y acción*, Universidad de Salamanca, Salamanca.

Broncano, F. (1994a), «La racionalidad de la ciencia y la naturalización de la epistemología», en *Actas del I Congreso de la Sociedad de Lógica, Metodología y Filosofía de la Ciencia*, UNED, Madrid.

Broncano, F. (1994b), «Prototypical Judgements and Skepticism about Rationality», en G. Munévar (ed.), *Philosophy of Science in the New Spain* (próxima publicación), Kluwer, Dordrecht.

Byrne, R. y Whiten A. (eds) (1988), *Machiavellian Intelligence. Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans,* Oxford University Press, Oxford.

Chance, M. R. A. y Mead, A. P. (1987), «Social Behaviour and Primate evolution», en R. Byrne y A. Whiten (eds.), 1987.

Cheney, D. y Seyfarth R. (1990), *How the Monkeys See the World. Inside the Mind of Another Species*, Chicago University Press, Chicago.

Cherniak, Ch. (1986), *Minimal Rationality* , MIT Press, Cambridge, Mass.

Clark, A. (1989), *Microcognition*, MIT Press, Cambridge, Mass.

Cosmides, L. (1989), «The Logic of Social Exchange: Has Natural Selection Shaped How Human Reason»: *Cognition, 31,* 187-276

Cosmides, L. y Toobby J. (1991), «From Evolution to Behavior», en E. Dupre (ed.), *The Latest on the Best*, MIT Pres, Cambridge, Mass.

Cohen, J. L. (1981), «Can Human Irrationality Be Experimentally Demonstrated?»: *The Behavioral and Brain Sciences, 4,* 317-370

Davidson, D. (1963), «Actions, Reasons and Causes», en *Essays on Actions and Events*, Clarendon Press, Oxford, 1980; original de 1960.

Davidson, D. (1970), «How is the Weaknes of the Will Possible?», en *Essays on Actions and Events*, Clarendon Press, Oxford, 1980, original de 1970.

Davidson, D. (1982), «Rational Animals»: *Dialéctica, 36,* 318-27

Davidson, D. (1992), «Engaño y división», en C. Moya (ed.), *Mente, Mundo y Acción,* Paidós, Barcelona.

De Sousa, R. (1987), *The Rationality of Emotions*, MIT Press, Cambridge, Mass.

Dennett, D. (1979), «Intentional Systems», en *Brainstorms*, Harvester Press, Hassocks.
Dennett, D. (1983), «Intentional Systems in cognitive Ethology: The Panglossian Paradigm Defended»: *Behavioral and Brain Sciences*, 6, 343-90.
Dennett, D. (1987), «Three Kinds of Intentional Psychology», en *The Intentional Stance*, MIT Press, Cambridge, Mass. (original de 1978). V.e.: Gedisa, Barcelona.
Dennett, D. (1992), *Consciousness Explained,* Little Brown, New York.
Ekman, P. (1992), «An Argument for Basic Emotions»: *Cognition and Emotion,* 6, 169-200
Elster, J. (1983) *Sour Grapes*. Cambridge University Press, Cambridge. V.e.: *Uvas amargas*, Península, Barcelona
Elster, J. (1989a), *Ulysses and the Syrens*, Cambridge University Press, Cambridge.
Elster, J. (1989b), *Solomonic Judgements. Studies in the Limitations of Rationality,* Cambridge University Press, Cambridge. V.e.: Gedisa, Barcelona.
Enç, B. y Adams, F. (1992), «Functions and Goal Directedness»: *Philosophy of Science, 59,* 635-654.
Gardenford, P. y Sahlin, N. (eds.) (1988), *Decision*, *Probability and Utility*, Cambridge University Press, Cambridge.
Ginet, C. (1990), *On Action*, Cambridge University Press, Cambridge.
Goldman (1970), *Human Action*, Princeton University Press, Princeton.
Gómez, J. C. (1989), «Visual Behavior as a Window for Reading the Mind of Others in Primates», en A. Whiten, 1989.
Gomila, A. (1993), «Evolución y racionalidad»: *Actas del I Congreso de la Sociedad de Lógica, Metodología y Filosofía de la Ciencia*, UNED, Madrid.
Goodall, J. (1989), *En la senda del hombre,* Salvat, Barcelona.
Gordon, R. M. (1987), *The Structure of Emotions*, Cambridge University Press, Cambridge.
Harris, P. (1992), *Los niños y las emociones,* Alianza, Madrid.
Hollis, M. (1987), *The Cunning of Reason,* Cambridge University Press, Cambridge.
Hollis, M. (1988), «The Social Function of Intellect», en Byrne y Whiten (eds.), 1988.
Jeffrey, R. C. (1965), *The Logic of Decision,* McGraw-Hill, New York.
Jolly, A. (1988), «Lemur Social Behavior and Primate Intelligence», en Byrne y Whiten (eds.), 1987.
Johnston, M. (1988), «Self-Deception and the Nature of Mind», en A. Rorty (ed.), *Perspectives on Self-Deception*, University of California Press, Los Angeles.
Johnson-Laird, P. M. (1983), *Mental Models*, Cambridge University Press, Cambridge.
Johnson Laird, P. M. y Oatley, K. (1992), «Basic Emotions and Folk Theory»: *Cognition and Emotion, 6,* 261-223
Kahneman, D., Slovic, P. y Tversky A. (eds.) (1982), *Judgement under Uncertainty. Heuristics and Biases*, Cambridge University Press, Cambridge.
Kalin, N. (1993), «Neurobiología del miedo»: *Investigación y Ciencia,* julio.

Le Doux, J. (1989), «Cognitive-Emotional Interactions in the Brain»: *Cognition and Emotion, 3*, 267-289.
Leslie, A. (1989), «The Theory of Mind. Impairment in Autism: Evidence for a Modular Mechanism of Development», en A. Whiten, 1989.
Luce, R. D. y Raiffa (1957), *Games and Decisions*, Dover, New York; reimpr. 1989, reimp. parcialmente en Moser, 1990.
Lyons, W. (1980), *Emotion*, Cambridge, University Press, Cambridge. V.e.: Anthropos, Barcelona.
Mele, A. (1987), *Irrationality. An Essay on Akrasia, Self-Deception and Self-Control*, Oxford University Press, Oxford.
Moser, P. K. (ed) (1990), *Rationality in Action. Contemporary Approaches*, Cambridge University Press, Cambridge.
Mosterin, J. (1978), *Racionalidad y Acción Humana*, Alianza, Madrid.
Mosterin, J. (1993), *Filosofía de la Cultura*, Alianza, Madrid.
Muguerza, J. (1977), *La Razón sin Esperanza*, Taurus, Madrid.
Muguerza, J. (1990), *Desde la perplejidad*, Fondo de Cultura Económica, Madrid.
Nisbett, R. y Ross, L. (1980), *Human Inference: Strategies and Shortcomings of Social Judgement*, Englewood Cliffs, Prentice-Hall.
Nozick, R. (1970), «Newcomb's Problem and Two Principles of Choice», en N. Rescher (ed.), *Essays in Honor of Carl G. Hempel*, Reidel, Dordrecht; reimp. en Moser, 1990.
Nozick, R. (1993), *The Nature of Rationality*, Princeton University Press, Princeton.
Parfitarfit, D. (1984), *Reasons and Persons*, Clarendon Press, Oxford.
Pears, D. (1984), *Motivated Irrationality*, Oxford University Press, Oxford.
Perner, J. (1989), *Understanding the Representational Mind*, MIT Press, Cambridge, Mass.
Premack, D. y Woodruf (1978), «Does the Cimpance Have a Theory of Mind?»: *Behavioral and Mind Sciences, 1*, 515-26
Putnam, H. (1988), «Racionalidad en la Teoría de la Decisión y en la Ética», en L. Olivé (comp.), *Ensayos sobre racionalidad en Ética y Política, Ciencia y Tecnología*, Fondo de Cultura Económica, México.
Quintanilla, M. A. (1981), «Fundamentos materialistas del racionalismo», en *A favor de la Razón*, Taurus, Madrid.
Railton, P. (1993), «Truth, Reason and the Regulation of Belief» (próxima publicación), en *Actas de la Conferencia SOFIA 1993, Verdad y Racionalidad*, Universidad de la Laguna, La Laguna.
Ramsey, F. P. (1931), «Truth and Probability», en *The Foundations of Mathematics and Other Logical Essays*, Routledge and Kegan Paul, London; rimp. en Gardenfords y Sahlin, 1988.
Rescher, N. (1993), *La racionalidad. Una indagación filosófica sobre la naturaleza y justificación de la Razón*, Tecnos, Madrid.
Ristau, C. A. (ed.) (1991), *Cognitive Ethology. The Mind of Other Animals. Essays in Honor of Donald R. Griffin*, Lawrence Erlbaum., Hilsdale, NJ.
Savage, L. J (^{2}1954), *The Foundations of Statistics*, John Wiley, New York; reimp. en Gardenfords y Sahlin, 1988.
Sen, A. y Williams, B. (1982), *Utilitarianism and Beyond*, Cambridge University Press, Cambridge.

Sloman, A. (1987), «Motives, Mechanisms and Emotions»: *Cognition and Emotion, 1*, 217-33.
Stich, S. (1990), *The Fragmentation of Reason*, MIT Press, Cambridge, Mass.
Tversky, A. y Kahneman D. (1982), « Judgements under Uncertainty: Heuristics and Biases», en Kahneman, Slovic y Tversky, 1982.
Whiten, A. (ed.) (1989), *Natural Theories of Mind*, MIT Press, Cambridge, Mass.

PERCEPCIÓN

Vicente Sanfélix Vidarte

I. INTRODUCCIÓN

La percepción constituye uno de esos asuntos perennes de discusión filosófica sobre el que sigue sin suscitarse consenso y generándose una extensa bibliografía. El objetivo de este trabajo es suministrar al lector un panorama de las principales posiciones que sobre el tema se han generado en la filosofía contemporánea, entendiendo por tal la que abarca nuestro siglo. He ordenado por ello la exposición bajo cuatro epígrafes, titulados respectivamente con los rótulos que he escogido para designar las cuatro perspectivas predominantes desde las que se ha enfocado, y sigue enfocándose, el problema, a saber: fenomenismo, realismo doxástico, representacionismo y teorías informacionales. Puesto que este trabajo se inscribe en un volumen dedicado a la mente, en la conclusión del mismo encontrará el lector una breve reflexión sobre alguna de las implicaciones que las distintas posiciones expuestas tienen para la concepción de lo mental.

Algunas advertencias más para terminar esta introducción. Si bien en cada epígrafe he mencionado una serie de nombres, atribuyéndoles la posición bosquejada en ellos, lo que ofrezco en cada caso no es sino una especie de retrato robot de la perspectiva aludida, del que no debe esperarse, por consiguiente, que haga total justicia a los puntos de vista de ningún autor particular. Por otra parte, aunque me he centrado en las teorías filosóficas de la percepción, intentando hacer ver más que nada lo que podríamos llamar su lógica interna: qué problemas las originan, a qué dificultades dan lugar, etc., como contrapunto he ido haciendo alusión a diferentes teorías psicológicas de la percepción. He optado por esta estrategia expositiva porque, sobre esta cuestión al menos, filosofía y psi-

cología nunca llegaron a divorciarse, por más que en el peor momento de sus relaciones —hacia la década de los cincuenta y de los sesenta— algunos filósofos y algunos psicólogos insistieran en la posibilidad de una demarcación tajante entre lo conceptual y lo empírico que debiera haber servido para separar claramente la filosofía de la psicología y, en general, de la ciencia. Notará el lector, por lo demás, que el contrapunto se pierde en el último epígrafe. La razón es que en este último epígrafe me hago eco fundamentalmente de las aportaciones al tema realizadas desde el ámbito de la ciencia cognitiva, una empresa que desde sus orígenes mismos se ha autoconcebido como interdisciplinar. Me he sentido, pues, libre de escrúpulos para hablar, sin separarlos, de filósofos y psicólogos.

II. TEORÍAS FENOMENISTAS

El fenomenismo, una posición que entronca con la tradición asociacionista que pasa por Stuart Mill en el XIX y se remonta hasta Hume y Berkeley en los siglos anteriores a éste, tuvo su apogeo en el primer tercio de nuestro siglo y suele asociarse con las posiciones defendidas en algunos de sus escritos por Moore y Russell, por Price o Ayer. Quizás quepa también calificar de fenomenistas algunas de las tesis de Husserl, si bien, inspirándose en sus últimos puntos de vista, Merleau-Ponty desarrolló una severa crítica del fenomenismo. Creo que pocos autores hoy defenderían el fenomenismo, aunque algunos, como Dicker, lo vindican parcialmente.

Antes que nada cabe considerar al fenomenismo como una respuesta al desafío escéptico que cuestiona la fiabilidad epistémica de nuestros sentidos. En efecto, parece de sentido común aceptar que percibir es una manera inmediata de obtener conocimiento de muchas cosas de las que los objetos físicos forman una clase destacada. Nos basta, por ejemplo, con echar un vistazo para estar absolutamente ciertos de que hay una mesa enfrente de nosotros.

Con esta creencia de sentido común ocurre, sin embargo, lo que con otras muchas, a saber: pronto la filosofía y también la ciencia parecen desafiarla. Pensemos si no en la multitud de ocasiones, que gustoso nos recuerda el escéptico, en las que por fiarnos de nuestros sentidos llegamos a abrazar creencias erróneas; desde los casos más cotidianos y pedestres, como aquellos en que confundimos un reclamo con un pato de verdad o tomamos por circular la torre de planta octogonal vista en la distancia, hasta los más sofisticados que nos descubren los psicólogos al presentarnos líneas de idéntica longitud dispuestas de tal modo que nos parecen claramente diferentes en este respecto. El gozo del escéptico crece cuando añade para nuestra consideración las experiencias alucinatorias y se desborda por último cuando nos presenta las contradicciones que parecen

darse entre nuestra imagen ordinaria del mundo y la que la ciencia nos presenta.

En efecto, desde el siglo XVII la física nos viene diciendo, por ejemplo, que no hay nada en las cosas semejante al color que creemos percibir en ellas. Y cuanto menos desde el XIX la psicofísica nos enseña que nuestra experiencia sensorial depende de manera esencial de la estructura de nuestro sistema nervioso. Un golpe, pongamos por caso, puede producir un destello si lo sufrimos en un ojo; pero un sonido, si se nos da en el oído. Fenómenos como éstos parecen legitimar la tesis de la energía específica de los diferentes nervios involucrados en cada una de las modalidades sensoriales, tesis que se asocia tradicionalmente al nombre del psicofísico J. Müller, y de la que parece deducible la conclusión escéptica de que las propiedades que percibimos, más que atribuírselas a los objetos del entorno, habría que cargarlas en la cuenta de nuestra constitución neuroanatómica.

Amalgamadas, todas estas consideraciones componen lo que se conoce como «el argumento de la ilusión», con sus versiones ilusiva, delusiva o causal. La tesis escéptica que del mismo puede sacarse parece clara. No debiéramos estar ciertos de que las cosas tienen las propiedades que parecen tener porque ahí están las ilusiones para recordarnos que podemos estar equivocados. No debiéramos estar ciertos ni tan siquiera de que las cosas que creemos percibir existen, porque ahí están las alucinaciones para recordarnos que también en este respecto podemos equivocarnos. Debiéramos concluir que lo que se ofrece a nuestra conciencia de una manera inmediata es el efecto del estado de nuestros órganos sensoriales y de nuestro cerebro —el estímulo próximo— y no los objetos situados en el espacio físico que puedan ser parcialmente responsables del mismo —el estímulo distante.

Hemos cuestionado ahora los puntos de vista del sentido común. Las creencias sobre el mundo físico que como resultado del ejercicio de nuestros órganos sensoriales obtenemos no merecen el título de conocimiento directo, esto es: inmediato e indubitable. Pero entonces, ¿hay algo que podamos conocer directamente gracias a nuestro aparato perceptivo? Y aunque no sea de una manera directa, ¿nos puede suministrar éste un conocimiento de los objetos que pueblan nuestro entorno?

En la segunda mitad del siglo pasado algunos psicólogos y neurofisiólogos de tradición empirista respondían a esta última pregunta afirmativamente. A partir de las sensaciones que experimentamos como resultado de la estimulación de nuestros órganos sensoriales, y mediante un proceso inferencial del que no tenemos consciencia, llegamos a percibir los objetos físicos que causan aquéllas. Son las líneas básicas del modelo que defendió Helmholtz.

Podemos preguntarnos, sin embargo, si esta explicación es coherente con el empirismo y si da cumplida cuenta del reto escéptico. Si afir-

mamos que de los objetos físicos sólo tenemos una conciencia que es el resultado de una inferencia causal, ¿no estamos concediendo que esos objetos son en sí mismos inobservables? Y si es así, ¿cómo podemos cerciorarnos de que son ellos efectivamente la causa de nuestras sensaciones? ¿No estamos convirtiendo a los objetos físicos en entidades teóricas a las que no podemos dotar de ningún contenido empírico? Más que un medio de acceder al mundo objetivo, las sensaciones parecen haberse convertido en una barrera o velo que nos impide entrar en contacto con él.

Como siempre, la solución más fácil es convertir la necesidad en virtud. Cierto que ninguna experiencia perceptiva puntual lleva en sí la garantía de su validez objetiva. Pero en toda experiencia, incluso en la alucinatoria, algo se presenta a la conciencia. Este algo a lo que la tradición le dio los nombres más variopintos —sensaciones, ideas sensibles, impresiones, representaciones, etc.— Moore lo bautizó con el término latino *sense-datum*, aunque fue Russell el primero que dejó constancia impresa de la nueva denominación.

Al escéptico le respondemos, para empezar, que en toda experiencia tenemos conocimiento directo: inmediato, completo e indubitable, de uno o varios datos sensoriales. Si quiere, le concederemos que podemos dudar de que estemos viendo realmente un pato, pero lo que no nos puede negar es que somos conscientes de unas manchas de color; de algo que es como la apariencia de un pato.

Los datos sensoriales son, pues, datos de consciencia. Como tales, tienen la característica de la privaticidad. Nadie comparte los datos sensoriales de otra conciencia, como nadie experimenta el dolor de muelas que atormenta al vecino. Por otra parte, son tan efímeros como la proyección de los fotogramas en la pantalla del cine; pues las manchas de color que ocupan nuestra consciencia visual cambian su tonalidad con la más leve alteración de la iluminación, pierden su forma original con sólo que movamos ligeramente nuestra cabeza, y a veces aun sin moverla en absoluto, como cuando decimos de lo que parece un pato que se ha movido.

Pero la respuesta al escéptico no es de momento sino parcial. ¿Podemos conocer los objetos físicos gracias a nuestra experiencia perceptiva, esa experiencia que ahora sabemos es de datos sensoriales? ¿Qué relación hay entre nuestra experiencia de éstos y nuestra conciencia de objetos? Consideremos el caso de la experiencia ilusiva. Experimentamos una minúscula mancha de color en nuestro campo visual, algo que tiene la apariencia de una torre circular en la distancia. A partir de esta conciencia podemos generar todo tipo de expectativas acerca de cuál será la naturaleza de nuestra experiencia visual en el futuro, acerca de qué datos sensoriales experimentaremos si, por ejemplo, dirigimos nuestros pasos en la dirección en que se encuentra eso que presenta la apariencia de una torre.

Pues bien, el caso es que algunas de esas expectativas se verán defraudadas. A medida que la mancha visual crece —síntoma de que nos vamos acercando a lo que creemos ser una torre— se nos aparece articulada en caras separadas por aristas. Salimos así de nuestro error. La torre, lo vemos ahora, es realmente de base octogonal.

La moraleja del ejemplo es ésta: la conciencia de objetos no es sino la conciencia de datos sensoriales que guardan entre sí ciertas relaciones. Somos víctimas de una ilusión cuando los datos sensoriales que percibimos en determinado momento nos llevan a esperar que los que experimentaremos en un futuro tendrán alguna propiedad o propiedades que de hecho no tienen. Si no tuvieran ninguna de las propiedades que esperábamos, más que de una ilusión de lo que habríamos sido víctimas es de una delusión; y ya no sería pertinente catalogar nuestra experiencia como perceptiva sino como alucinatoria.

La conciencia de objetos no es, pues, inmediata. En ella hay involucrada una inferencia, aunque no causal sino inductiva. Percibir un objeto físico no es inferir una causa inobservable para un dato inmediatamente experimentado, sino inferir que este dato inmediatamente experimentado será seguido por otros de ciertas características. La experiencia, que es un conjunto de estados sucesivos constituidos cada uno de ellos por la totalidad de datos sensoriales que se experimentan coetáneamente, tendrá validez objetiva cuando sea coherente, esto es: cuando esos datos se confirmen mutuamente. Y carecerá de ella en caso contrario. ¿Basta esto para acallar al escéptico?

III. REALISMO DOXÁSTICO

Los filósofos que podemos situar en esta nueva posición: Wittgenstein, Ryle, Austin, Quinton, Armstrong o Pitcher, entre otros, consideraban que la respuesta a esta última pregunta era negativa.

¿Cuántos datos sensoriales confirmatorios deberemos reunir para disipar la sospecha de que nuestra experiencia sea ilusoria o alucinatoria? La pregunta no tiene una respuesta definida. Con cada objeto físico asociaremos siempre sólo un número finito de datos sensoriales efectivos, pero debemos reconocer, si optamos por el fenomenismo, que el número de datos sensoriales que podríamos relacionar con ese mismo objeto físico es potencialmente infinito. Desde cada punto del espacio lo que parece ser una torre ofrecería un aspecto diferente, y el dato sensorial que se asocia a cada punto de ese espacio debemos entender que se disipa y se recrea casi con los instantes mismos del tiempo. Puesto que el objeto físico hay que entenderlo como el conjunto tanto de estos datos sensoriales posibles cuanto de aquellos que efectivamente experimentamos, y dado que los primeros siempre exceden en número infinito a los

segundos, postular aquéllos a partir de la constatación de éstos parece una temeridad desde un punto de vista lógico. Dicho en otros términos: la base empírica a partir de la cual inducimos la existencia de los objetos físicos parece condenada a ser siempre excesivamente pobre. Lo sensato sería concluir que nuestra creencia inducida en la existencia de cualquier objeto físico es siempre falsable, pero nunca exhaustiva o definitivamente verificable. Los objetos físicos se han vuelto a convertir en entidades teóricas cuya existencia resulta sumamente problemático justificar empíricamente. Por otra parte, desde una óptica fenomenista resulta difícil explicar cómo pueda aprenderse el lenguaje y, por extensión, el fenómeno mismo de la comunicación lingüística. Si cuando quiero enseñar a alguien el significado de una palabra o de una oración, la referencia de aquélla o las condiciones de verdad de ésta no las constituyeran sino mis datos sensoriales, difícilmente podría el aprendiz establecer la conexión pertinente entre mis expresiones y sus condiciones semánticas. Pues mis datos sensoriales son algo que, por principio, él no podría experimentar.

Se pueden, además, bloquear los argumentos del escéptico sin tener que recurrir para ello a la noción de dato sensorial. A quien aduzca el tan traído y llevado argumento de la ilusión, cabrá recordarle que, como ya dijimos, este argumento se apoya en consideraciones científicas. Pues bien, la ontología que la ciencia asume es perfectamente objetiva. El psicólogo, por ejemplo, se puede permitir muchas dudas, pero no la de que es una y la misma ilustración de la ilusión de Müller-Lyer la que resulta visible para los sujetos experimentales a los que se la suministra, pues de lo contrario nada de su actividad experimental tendría sentido. El escéptico sólo puede formular sus dudas sobre la validez objetiva de nuestras creencias perceptivas aceptando, como hace la ciencia en cuyos argumentos se apoya, la fiabilidad de las mismas.

Y si insiste, amparado en los casos en que nuestros sentidos nos inducen a adoptar creencias erróneas, en desacreditar la fiabilidad epistémica de los mismos, siempre podremos esgrimir en su contra que no tiene sentido hablar de experiencias erróneas si no podemos contraponerlas a las experiencias verídicas. Aquéllas parecen exigir éstas como el falsificador necesita de las monedas de curso legal para poder transgredir la ley. Y no debemos volvernos especialmente desconfiados hacia nuestra experiencia, no sólo porque los errores y las alucinaciones, como las falsificaciones, son después de todo detectables, sino sobre todo porque aquéllas, como éstas, tienen que ser necesariamente la excepción antes que la regla; pues del mismo modo en que una actividad falsificadora irrestricta, cuyo resultado sería una sobreabundancia de dinero, generaría una situación de inflación que perjudicaría por igual al honrado y al criminal, una situación de ilusiones y alucinaciones continua arruinaría el valor significativo de esas monedas de la comunicación que son las pa-

labras, lo que afectaría por igual al sentido de lo que dice el crédulo como al de lo que dice el escéptico.

Lo que acaba de decirse quizás se entienda mejor si reparamos en que entre la evidencia objetiva y pública y el significado de nuestras palabras hay una conexión que no puede romperse gratuitamente. Estamos situados a dos metros de la mesa de nuestro comedor. La iluminación es adecuada. ¿Por qué habríamos de dudar de que estamos viendo una mesa? Enseñamos o aprendemos el significado de los verbos de percepción y de las expresiones para referirnos a los objetos percibidos en condiciones semejantes a éstas. Si alguien, como el escéptico, pretende que en todas las situaciones de este tipo cabe la duda, está cuestionando la condición que hace posible que las palabras, las suyas incluidas, tengan significado. Por eso en condiciones normales no tengo por qué justificar mi creencia de que estoy viendo una mesa, pues es el significado de la palabra «mesa» el que permite que en situaciones como éstas resulte legítimo decir que enfrente de nosotros hay una mesa, y es el escéptico el que debe justificar su duda dándonos razones para sospechar que ésta no es una situación normal.

Vemos ahora lo estéril de la estrategia del fenomenista. Se ha empeñado en encontrar enunciados que en virtud de sus referentes llevaran en sí una garantía de certeza. No se daba cuenta de que toda descripción, por suponer la clasificación de las propiedades de la entidad descrita bajo uno o varios predicados, es susceptible de error. Pues clasificar es, como cualquier otra actividad, algo que podemos hacer mal. Pero ello no debe llevarnos a la desesperación escéptica, porque la certeza que ningún enunciado empírico es capaz de suministrarnos por sí mismo, nos la pueden aportar las circunstancias en que lo aplicamos. Al fenomenista hay que reprocharle que estuviera buscando la certeza en la dirección errónea: en los referentes de los enunciados en vez de en las condiciones de utilización de los mismos. Quizás por esa desorientación original ha puesto luego todo del revés.

Si, como hemos dicho, una condición del aprendizaje del lenguaje es que lo utilicemos para referirnos a objetos públicos e intersubjetivos, de ello se sigue que sólo cuando dominemos los predicados para referirnos a estos objetos podremos utilizarlos para caracterizar nuestra experiencia de los mismos. Tenemos que empezar por saber cuándo toca decir de algo objetivo, la leche por ejemplo, que es blanca, para poder decir después que experimentamos visualmente algo de aspecto blanco. Describimos nuestra experiencia utilizando términos que tomamos prestados de nuestra descripción del mundo. Nuestra descripción, y consiguientemente nuestra certeza, de las propiedades de nuestra experiencia es parasitaria de la descripción y de la certeza que tenemos respecto a las propiedades de las cosas externas. Si estamos aprendiendo italiano y sospechamos de «rosso» que es el término para un color, aunque no sa-

bemos exactamente cuál, tanto dudaremos de describir la sangre como «rossa», cuanto de decir, cuando la estamos mirando, que nuestra experiencia es la de algo que presenta un aspecto «rosso».

Disipada la prioridad lógica del lenguaje de los datos sensoriales sobre el lenguaje objetivo, el privilegio epistemológico de aquéllos sobre los objetos físicos se disuelve como los azucarillos en el agua. Cierto que cuando utilizamos el modismo de las apariencias —«me parece que...», «me parece como si...», etc.— lo que decimos resultará la mayor parte de las veces incorregible, pero no debemos olvidarnos de que tales enunciados tanto pueden versar sobre objetos físicos y sus propiedades, cuanto sobre nuestra experiencia de los mismos, y la incorregibilidad ganada no se debe al tipo de entidades al que nos referimos, sino a la manera cautelosa y poco comprometida que hemos escogido para hablar de ellas. El sentido común parece ahora reafirmado. Nuestros actos objetivos parecen tener como sus destinatarios inmediatos los objetos de nuestro entorno físico. Por eso los psicólogos de la Gestalt parece que actuaron sensatamente cuando reaccionaron contra los psicólogos y fisiólogos empiristas y su parafernalia de inferencias o asociaciones. Como ellos, estamos ahora en disposición de defender que nuestra experiencia es, de manera espontánea e inmediata, una experiencia de realidades objetivas, y es esta objetividad la que explica la naturaleza de nuestra experiencia y no al revés; como podemos concluir reflexionando sobre los fenómenos de percepción de figuras ambiguas como el famoso pato-conejo de Jastrow, la vieja y la dama de Boring, etc.

Se trata de casos en los que la organización de la experiencia se altera automáticamente conforme varía nuestra interpretación del significado del dibujo. Así, lo que antes veíamos como unas orejas, cuando interpretábamos el dibujo como el de un conejo, lo vemos, ahora que el dibujo nos parece el de un pato, como el pico del animal.

El significado de estos fenómenos no es sin embargo claro. Se pueden hacer valer desde luego contra el fenomenismo y a favor de una concepción «gestáltica» de la experiencia perceptiva, si simplemente nos limitamos a señalar que muestran que la organización de esa experiencia no es previa ni independiente al valor objetivo que le atribuimos. No hay uno y siempre el mismo dato sensorial a partir del cual inferimos unas veces la cabeza del conejo y otras la del pato por la sencilla razón de que el contenido de nuestra experiencia, su organización general, su forma o Gestalt, varían de golpe según interpretemos los dibujos. Pero también pueden hacerse valer estos fenómenos para ir más allá de una concepción gestáltica de la percepción si insistimos en el carácter contextual que tienen nuestras interpretaciones del dibujo.

En efecto, los psicólogos de la Gestalt pensaban que la organización de la experiencia sensorial estaba endógenamente constreñida por condicionamientos de tipo neurofisiológico. Pero resulta más fácil ver otro

tipo de constricciones que operan sobre nuestra interpretación de la experiencia, tales como nuestra competencia conceptual; pues difícilmente podrá ver el dibujo del pato-conejo como un pato quien no haya visto jamás ninguno, ni tenga un concepto formado respecto a los mismos.

Los filósofos a los que hemos aludido insistían en este carácter conceptual e inteligente de las actividades perceptivas. Para ellos la percepción no involucraba procesos inferenciales, y este reproche contra el fenomenismo los acercaba a las posiciones de los psicólogos de la Gestalt. Pero también para ellos la inmediatez había que relativizarla al equipamiento conceptual del sujeto perceptor. Sólo si tengo los conceptos de pato y de conejo puedo ver de manera inmediata la figura de Jastrow como una representación de cualquiera de los dos. En consecuencia, consideraban la percepción como una actividad esencialmente inteligente por la que el sujeto perceptor, sirviéndose de sus órganos sensoriales y condicionado por su esquema conceptual, obtenía de una manera inmediata creencias sobre el entorno. Es la insistencia en este elemento doxástico lo que aleja sus tesis de las de los psicólogos de la Gestalt, y lo que les acercaba tanto a las posiciones de los psicólogos transaccionalistas, y en general de los partidarios del *new look*, cuanto a los conductistas. Como los primeros, asumirían la permeabilidad de la experiencia perceptiva a la influencia de otros procesos psíquicos de naturaleza cognitiva e incluso emocional; como los segundos, insistían en el carácter aprendido de buena parte de nuestras habilidades perceptivas dado que los conceptos que las posibilitan no son innatos, en la naturaleza parasitaria de nuestro conocimiento y descripción de la experiencia, o en la naturaleza disposicional de las creencias en que la actividad perceptiva cristaliza.

IV. REPRESENTACIONISMO

Si el realismo doxástico cabe entenderlo como una reacción frente al fenomenismo, la revitalización del representacionismo o de las teorías causales que se ha venido produciendo en la segunda mitad de nuestro siglo puede ponerse en relación con las dificultades de aquél. Autores como Grice, Goldman, Mackie o Jackson son de obligada referencia cuando se trata de entender las líneas maestras de esta posición.

Aun si podemos considerar que el realismo doxástico da cumplida cuenta del argumento de la ilusión y de otras estratagemas escépticas que quieren hacernos dudar de que el mundo objetivo sea el destinatario de nuestras actividades perceptivas, no está claro, sin embargo, que no abra las puertas a ese hermano menor del escepticismo que es el relativismo, porque si ese proceso de adquirir creencias inmediatas acerca del mundo que es percibir es siempre relativo a un esquema concep-

tual, la pregunta que inmediatamente se plantea es: ¿qué ocurre cuando dos sujetos se enfrentan a una misma situación equipados con conceptos o teorías diferentes?

Hanson se planteó una concreción de este problema. ¿Verían lo mismo un astrónomo medieval y uno contemporáneo al otear la bóveda celeste? Siguiendo el hilo conductor de las reflexiones wittgensteinianas sobre la percepción de figuras ambiguas, propuso una respuesta negativa a la cuestión, lo que no hace sino confirmar la sospecha recién esbozada. Por otra parte, se ha denunciado también el relativismo como una consecuencia inevitable de alguno de los planteamientos psicológicos —los del *new look*, fundamentalmente— con los que el realismo doxástico guardaría, según vimos, ciertas afinidades.

Para hacer de ello motivo de crítica tendríamos que demostrar, no obstante, que el relativismo perceptivo es una tesis epistémicamente indeseable y empíricamente injustificable, cuestión cuanto menos discutible; y lo cierto es que el realismo doxástico parece que puede cuestionarse de una manera más directa por no atrapar ni las condiciones necesarias ni las suficientes de los fenómenos perceptivos.

En efecto, habíamos concluido que percibir era una peculiar manera —a través de los sentidos— de obtener creencias. Empero, en determinadas ocasiones estamos dispuestos a predicar de algunos sujetos actividades perceptivas sin por ello atribuirles ninguna creencia como resultado de las mismas. Es el caso paradigmático de los animales inferiores y de los bebés humanos, o incluso de los humanos adultos cuando éstos no prestan una especial atención a lo que están percibiendo. Lo que estos casos de percepción no epistémica —como la denominó Dretske— parecen demostrar es lo innecesario del componente doxástico como elemento de los procesos perceptivos. Sencillamente, no siempre quien percibe forma creencias acerca de lo que percibe; y en realidad parece intuitivo afirmar que siempre el ámbito de lo percibido excede en riqueza el ámbito de lo creído por mor de la percepción.

Por otra parte, tampoco parece suficiente tener una creencia verdadera sobre algo, obtenida gracias al ejercicio de los sentidos, para que estemos legitimados a hablar de un genuino caso de percepción de ese algo. Imaginemos que se nos pide, en determinada situación experimental, que digamos de una vela, situada a unos metros de distancia enfrente de nosotros, si está encendida o apagada, y que las condiciones experimentales son tales que, interpuesto entre esa vela —llamémosla *a*— y nosotros, hay un espejo cuya existencia desconocemos y que refleja otra vela —llamémosla *b*—. Por lo demás, ambas velas, *a* y *b*, son siempre o encendidas o apagadas a la vez. Así pues, en esta situación nos formaremos, gracias al ejercicio de nuestros sentidos, una creencia verdadera acerca del estado de *a*, aunque en realidad no es *a* sino *b* la vela que estamos viendo.

Probablemente el realista doxástico pueda refinar sus propuestas hasta dar cabida a este contraejemplo. Pero cuando la consideración del mismo se suma a las consideraciones anteriores acerca de lo innecesario del componente doxástico en los procesos perceptivos, la conclusión que parece imponerse es que el realista ha descuidado más de la cuenta un factor que parece consustancial a éstos: la experiencia sensorial.

Al bebé le atribuimos capacidades perceptivas, aunque somos reticentes a atribuirle creencias, porque a menos de que dispongamos de evidencia en sentido contrario, evidencia neurofisiológica o conductual, lo suponemos sujeto de esa experiencia. Del mismo modo, podemos decir que no nos hemos fijado, no que no hayamos visto, en si quien nos asaltó llevaba o no bigote, porque aunque el miedo nos impidió prestar atención a este detalle, suponemos que nuestra experiencia visual hubiera sido significativamente diferente de estar o no la cara de nuestro asaltante adornada por un mostacho. Y a estas consideraciones, basadas en nuestras prácticas lingüísticas ordinarias, cabría añadir la mucha evidencia experimental que apunta en el sentido de que buena parte de la información suministrada por los órganos sensoriales es desatendida en fases del proceso perceptivo demasiado tempranas como para que el sujeto pueda llegar a formar ninguna creencia sobre ella. Añadamos ahora a todo ello que en el experimento mental recién descrito lo que nos induce a decir que lo percibido es la vela *b* y no la *a* es que la primera, y no la segunda, es la responsable causal de la experiencia visual del sujeto experimental, y el caso contra el realismo doxástico parecerá visto para una sentencia condenatoria.

Precisamente lo que los teóricos representacionistas han hecho es tomarse completamente en serio la experiencia sensorial hasta convertirla en un elemento fundamental del análisis de los procesos perceptivos. Lo común a todas las propuestas representacionistas es la insistencia en la necesidad de introducir en ese análisis una cláusula que especifique el nexo causal entre la experiencia del sujeto perceptor y el objeto percibido. De esta manera, tales propuestas sintonizan con la revitalización del mentalismo que, tras un largo periodo de austeridad conductista, viene produciéndose en el ámbito de la psicología desde finales de la década de los sesenta.

En efecto, desde un modelo conductista radical como, por ejemplo, el skinneriano, la percepción debía explicarse, como cualquier otro fenómeno psíquico, en términos de estímulos y respuestas, variables ambas directa e inmediatamente observables. Se desdeñaba como irrelevante, cuando no como metodológicamente perversa, la postulación de cualquier variable intermedia, incluso si ésta se interpretaba en términos neurofisiológicos. Explicar las habilidades perceptivas de un organismo era enumerar cómo y a qué estímulos era capaz de reaccionar.

El mensaje metodológico del representacionismo es que esto no basta. Que si queremos explicar lo que es percibir, necesariamente hemos de atender a las propiedades de un elemento que se intercala entre estímulo y respuesta: la experiencia. Ahora bien, la cuestión que inmediatamente se plantea es si con la reintroducción de la experiencia como eslabón intermedio entre el mundo y nuestras creencias perceptivas o, más en general, nuestra conducta discriminativa, no se echan por la borda las conquistas anti-escépticas tan duramente ganadas por el realismo.

Si una condición de percibir algo es que ese algo sea causalmente responsable de nuestra experiencia sensorial, podría pensarse que justificar, contra la duda del escéptico, que efectivamente estamos percibiéndolo exige demostrar que entre el objeto supuestamente percibido y nuestra experiencia sensorial existe un efectivo nexo causal. Ahora bien, ¿cómo podríamos hacer esto si el objeto no nos es accesible sino a través de la experiencia sensorial? He aquí planteado de nuevo el problema del velo de la percepción.

Algunos representacionistas, como Mackie o Jackson, han pensado que la única salida del atolladero consiste en conceder que nuestra creencia en la validez objetiva de nuestros procesos perceptivos sólo puede justificarse mediante una especie de inferencia de la mejor explicación. Del mismo modo en que, aunque no podamos percibirlas, creemos en la existencia de partículas sub-atómicas porque su postulación nos da la mejor explicación de los fenómenos físicos que podemos observar, debemos creer en la existencia de un mundo objetivo porque su postulación nos suministra la mejor explicación de las propiedades de la experiencia de la que somos conscientes.

No encuentro muy confortativa esta réplica al escéptico, que a mi entender condena a los objetos físicos al status de entidades teóricas inobservables. Ni tampoco la encuentro, a decir verdad, necesaria; pues no veo por qué el representacionista no puede disponer, en principio, de los mismos argumentos trascendentales que contra las dudas pirrónicas esgrimió el realista doxástico. Se trata de recordarle al escéptico, una vez más, que sus propias dudas no pueden formularse inteligiblemente si no supone un mundo público e intersubjetivo al que nos dan acceso nuestros órganos sensoriales. Pero ahora se añade que parte de lo que esto significa es admitir que la experiencia sensorial, la suya no menos que la nuestra, es causalmente elicitada por las entidades de ese mundo objetivo. Si el realista concluía de estos argumentos la validez objetiva de nuestras creencias perceptivas, el representacionista concluirá la validez objetiva de la experiencia sensorial en la que éstas se basan. No veo por qué haya de ser más problemática esta conclusión que aquélla .

Por otra parte, el representacionista puede reforzar con un argumento de tipo naturalista la fiabilidad de nuestra experiencia sensorial. Al

fin y al cabo también gracias a ella hemos pasado la prueba de la selección natural, lo que es una buena razón para pensar que la información del entorno que nos suministra es lo suficientemente fiable como para permitirnos desarrollar una conducta adaptativa.

V. TEORÍAS INFORMACIONALES

Acabamos de mencionar la palabra «información». Este es el término clave que califica a la práctica totalidad de teorías, no siempre obviamente compatibles, que se proponen en la actualidad para dar cuenta de la naturaleza de los procesos perceptivos, tanto en el ámbito filosófico como en el psicológico o en el de la inteligencia artificial. De informacionales cabe tildar las teorías de Gregory o Rock, las de Marr, las tesis de Fodor, las de Dretske y también las de Gibson.

De estos autores, la mayoría se mueven en el marco general del cognitivismo, concibiendo la mente según el modelo del computador y la percepción como un fenómeno de procesamiento de la información. Quizás se entienda mejor lo que esto significa si reflexionamos, siquiera sea brevemente, sobre el proceso de percepción visual.

Parece difícilmente discutible que este proceso se inicia cuando la luz refractada por los objetos del entorno incide sobre nuestros ojos abiertos y causa en nuestras retinas un determinado patrón estimular. Obviamente, este patrón no es aleatorio, sino que está en función de la naturaleza de la luz que, en virtud de sus propiedades físicas, han refractado los objetos. Según sea, por ejemplo, la longitud de onda de la luz refractada, se activarán diferencialmente en la retina los tres tipos de conos de que ésta dispone para la detección de esta propiedad lumínica.

Ahora bien, puesto que cuando entre dos eventos se da una relación nómica, decimos que uno de ellos lleva información sobre el otro; y puesto que cuando alguien es capaz de servirse de esta información para, a partir de la constatación de uno de esos eventos, llegar a determinar el otro decimos que para ese alguien el primer evento representa el segundo, bien podríamos decir que los patrones estimulares retinianos, en virtud de su relación legaliforme con los objetos situados en el entorno que refractan la luz incidente en nuestros ojos, son representaciones que llevan información sobre éstos, pues al sujeto perceptor estos patrones le permiten llegar a conocer determinadas propiedades de esos objetos.

Sin embargo, aunque los patrones estimulares retinianos son para nosotros representaciones de los objetos del entorno, nosotros no percibimos ni somos conscientes de los mismos. Basta para tener que aceptar esta conclusión considerar que, si bien las imágenes retinianas son dobles, una para cada ojo, la escena visual de la que tenemos consciencia es, en la inmensa mayoría de los casos, única.

De aquí se sigue una tesis de importancia vital, a saber: que la percepción es un proceso en el que están involucradas representaciones no todas las cuales tienen por qué ser conscientes o, como también se dice a veces, fenomenológicamente accesibles al sujeto perceptor. En el caso de la percepción visual, partiendo de esas dos representaciones inconscientes que, según hemos visto, serían los patrones estimulares retinianos, llegamos, mediante algún tipo de proceso, a una única representación, la que se presenta a nuestra consciencia, de la escena visual. Es obvio que de este proceso tenemos tan poca noticia introspectiva como de muchas de las representaciones sobre las que opera, pero ¿podemos precisar algo más de su naturaleza?

Si tenemos en cuenta que la retina y, por consiguiente, los patrones estimulares que en ella se forman, son bidimensionales, y que sin embargo la escena visual de la que llegamos a ser conscientes es tridimensional, podemos comprender el sentido que tiene decir de estos procesos que son enriquecedores: el punto de partida, los estímulos próximos, son más pobres que la distribución del estímulo distante finalmente percibido.

El fenómeno de las constancias, sobre el que tanto insistieran los psicólogos de la Gestalt, sirve para ilustrarnos otra importante característica de estos procesos: su naturaleza inteligente, entendiendo por tal el proceder de conformidad con determinadas reglas. Por ejemplo, a pesar de que el patrón estimular que un objeto que se acerca hacia nosotros provoca en nuestra retina varía continuamente, nosotros seguimos percibiendo el tamaño de ese objeto como constante. Curiosamente fue Koffka, uno de los más destacados representantes de la escuela de la Gestalt, quien advirtió que en este fenómeno hay implicada una relación invariante entre el ángulo subtendido por el objeto desde nuestro ojo y la distancia a la que el objeto es percibido; y digo «curiosamente» porque una explicación que parece natural de este fenómeno de constancia, una vez que este detalle del mismo se ha puesto de relieve, sería, contra el espíritu de la teoría gestáltica, afirmar que determinamos el tamaño de los objetos que se mueven con respecto a nosotros calculando inconscientemente la relación entre las dos variables recién aludidas, de tal manera que cuando el resultado del cómputo es invariante, se percibe el objeto como de un tamaño constante, independientemente de cómo varíe la magnitud de la región que ocupa en nuestro campo visual.

Lo que tenemos aquí no es ni más ni menos que una reivindicación del viejo modelo Helmholtziano. Podemos dar una formulación general del mismo en los siguientes términos: la percepción no es sino un proceso por el que, partiendo de representaciones inconscientes de estímulos próximos, llegamos, mediante inferencias enriquecedoras o inductivas igualmente inconscientes, a una representación de la configuración del estímulo distante causalmente responsable de aquellas primeras representaciones.

Sin embargo, muchos de los psicólogos defensores de este modelo insisten ahora en que el mismo puede limpiarse de toda excrecencia empirista. Los principios computacionales o inferenciales a que obedecen los procesos perceptivos no son aprendidos sino innatos, por ello mismo tampoco está a nuestro alcance alterar a voluntad el funcionamiento de los mismos. Ello significa que nuestra experiencia perceptiva es impermeable a la influencia de otros elementos psicológicos como las creencias o los deseos; y, como prueba, los partidarios de este enfoque insisten en la inevitabilidad de las percepciones ilusorias que no conseguimos evitar ni aun cuando se nos avisa de su carácter.

En efecto, de nada nos sirve que se nos advierta, cuando por ejemplo somos víctimas de la ilusión de Müller-Lyer, de que las líneas dibujadas en el papel son de igual longitud, pues nosotros seguimos viendo la una como mayor que la otra. Y aunque es cierto que la organización de las figuras ambiguas puede variar, lo cierto es que la plasticidad de esa organización tiene límites: podemos ver la imagen de Boring como el dibujo de una anciana o de una joven muchacha, pero no como el de cualquier otra cosa.

La impermeabilidad de la experiencia sensorial no es sino una consecuencia de la concepción modular de los procesos perceptivos. Éstos son concebidos como funcionando autónomamente, como módulos, de los que sólo su resultado entra en relación con otros mecanismos psicológicos. Las implicaciones epistemológicas de este enfoque son claras. Si como vimos era posible que el realismo doxástico abriera las puertas del relativismo, la concepción de la percepción como un proceso modular parece en disposición de cerrar el paso a este vástago del escepticismo. Pero quizás la mayor significación filosófica de este tipo de enfoques sea otra, a saber: la posibilidad que abren de que se disponga de una explicación naturalizada de los procesos psíquicos y, consiguientemente, de legitimar de una manera rotunda la clasificación de la psicología entre las ciencias naturales.

En efecto, tradicionalmente se ha visto el carácter intencional de los estados mentales, su versar sobre algo, como un motivo para reivindicar un status científico peculiar para la psicología. Pero lo que las teorías cognitivas nos permitirían ver es que la dimensión intencional no impide, sino que presupone, la existencia de relaciones nómicas entre los estados mentales y el mundo físico, pues es la existencia de estas relaciones la que da cuenta del carácter informacional que convierte a los mismos en fenómenos representacionales o intencionales.

Pero no todos los teóricos que han hecho de la información el concepto clave de sus teorías de la percepción han reivindicado el modelo helmholtziano que acabamos de esbozar. Es el caso de los partidarios del llamado «enfoque ecológico», con J. J. Gibson como su principal mentor.

Cuando, por ejemplo, los teóricos ecológicos explican la percepción

visual, también parten del principio de que la estructura de la luz refractada lleva información sobre el entorno, pero no entienden éste como una categoría física sino ecológica, lo que quiere decir que lo que cuenta como entorno no puede definirse al margen del tipo de organismo que lo habita. De esta forma, lo que caracterizado en términos físicos contaría como un único entorno puede ser, desde un punto de vista ecológico, la realización de entornos muy diferentes. Si, por consiguiente, queremos conocer el tipo de información que lleva la luz, no nos servirá la óptica física ni la geométrica, sino que deberemos relativizar la pregunta a un tipo de organismo, deberemos enfocar la cuestión como un asunto propio de la óptica ecológica.

Las diferencias con los cognitivistas son, a partir de aquí, radicales. Como hemos visto, el punto de partida de los procesos perceptivos son para éstos una representación empobrecida del entorno que va paulatinamente enriqueciéndose hasta originar una representación final del estímulo distante, del que, en el caso de la percepción visual, alguno de los teóricos cognitivos más destacados ha opinado que no nos suministra sino una presentación de sus propiedades geométricas y cromáticas, base informativa a partir de la cual otros procesos psíquicos, ya no estrictamente perceptivos sino de naturaleza más intelectual, podrían lograr el reconocimiento categorial, funcional o semántico de los objetos percibidos, esto es: de qué tipo son, para qué sirven o cuál es su significado.

Pues bien, los teóricos ecológicos niegan el punto de partida mismo de las teorías cognitivas: el argumento de la pobreza estimular; y negando esto niegan también que percibir sea un proceso inferencial de enriquecimiento paulatino de representaciones. Para ellos la percepción es directa, una actividad exploratoria por la que el organismo extrae, que no procesa, la información contenida en la luz. Es el sujeto perceptor quien sirviéndose de sus sistemas perceptivos, que en el caso de la visión no incluyen sólo los ojos, más el nervio óptico, más el cerebro, sino también la cabeza móvil en que todos aquéllos están colocados y aún el cuerpo igualmente móvil en que ella se inserta, puede captar las invariantes que la experiencia ofrece, invariantes que no lo informan de propiedades geométricas de los objetos del entorno sino, desde el principio, de las utilidades que éstos tienen para él. Así, por ejemplo, lo que un animal percibe cuando se le sitúa sobre una sólida superficie de cristal transparente, situada a cierta altura del suelo, no serían las propiedades geométricas de la profundidad o de la tridimensionalidad del espacio, sino la realidad ecológica de un abismo, un lugar propicio para la caída, en suma: la utilidad negativa o nula que tal superficie proporciona para la locomoción; de ahí su deslizarse cautamente a través de ella o su adoptar la posición de caída.

La teoría ecológica parece implicar una vuelta a las coordenadas del realismo. Muy probablemente su insistencia en la dimensión práxica

de la actividad perceptiva la sitúe en una posición mejor que la de los teóricos doxásticos para dar cuenta de los casos de percepción no epistémica, y quizás tampoco resultara muy difícil acomodar en ella el ingrediente causal que parece intuitivamente un elemento ineludible de nuestras concepciones pre-teoréticas de los fenómenos perceptivos. Pero si tenemos en cuenta que el entorno de los seres humanos no es sólo ecológico sino también socio-cultural, cabría preguntarse si también ella implica, como algunos han sospechado, cierto grado de relativismo.

VI. CONCLUSIÓN

El lector habrá notado que el hilo conductor que ha guiado la exposición de las diversas teorías de la percepción ha sido fundamentalmente epistemológico. No es casualidad, pues creo que los filósofos se han preocupado del problema de la percepción básicamente desde esta perspectiva. La cuestión más perentoria para ellos ha sido si la percepción es una fuente fiable de conocimiento; o dando este punto por asumido, cómo hay que entender la percepción para explicar que sea una fuente adecuada de conocimiento.

¿Qué repercusiones tienen entonces las teorías de la percepción expuestas para la filosofía de la mente? Yo diría que para el problema mente-cuerpo podríamos considerar muchas de ellas, si no como estrictamente neutrales, sí al menos como más o menos acomodables a diversas soluciones del mismo. Por poner un ejemplo especialmente radical del asunto: ser un dualista impenitente o un materialista contumaz no impide suscribir una teoría representacionista.

Sin embargo, creo que, más allá del problema mente-cuerpo, las teorías de la percepción sí que tienen implicaciones de grueso calibre para la concepción de lo mental, en particular: para la concepción del sujeto psicológico. Más en concreto diré que el fenomenismo, el representacionismo y las teorías cognitivas comparten un determinado concepto de la experiencia perceptiva y, consecuentemente, del sujeto de la misma; un concepto por el que aquélla queda categorizada como una agregación de elementos separables, coetánea o sucesivamente experimentados, y aquél, expresándolo en términos cibernéticos, como un sistema de estados discretos.

Quizás lo que quiero decir quede más claro si afirmo que esta concepción de la vida psíquica y del sujeto de la misma sería el resultado de haber entendido ambos según el modelo de la física matemática. Ryle habló de cartesianismo y de concepción para-mecánica de lo mental. Quizás debiera haber hecho referencia a Hume y su propuesta de entender el sujeto psíquico como un microcosmos newtoniano.

Probablemente implicada por esta concepción de la experiencia per-

ceptiva y del sujeto de la misma esté la tesis de que una entidad sin capacidad de intervención en el mundo, pero que estuviera estimularmente relacionado con él, podría en principio merecer el título de sujeto perceptor. Quizás algo como el alma desencarnada del cartesianismo o su versión contemporánea: el cerebro en la cubeta.

A los realistas, ya sean doxásticos o ecológicos, esta conclusión les parecerá una reducción al absurdo de cualquier tesis de la que se siga. Pues sin acción no hay intencionalidad ni, por consiguiente, percepción. Frente a la concepción atomizada de la experiencia y la correspondiente concepción modular del sujeto de la misma, los realistas opondrían una concepción continua de la primera y básicamente unitaria del segundo. El sujeto perceptor es siempre el organismo, o ese organismo enculturado y social que es la persona.

BIBLIOGRAFÍA

Armstrong, D. M. (1961), *Perception and the Physical World*, Routledge and Kegan Paul, London. V.e.: Tecnos.

Ayer, A. J. (1940), *The Foundations of Empirical Knowledge*, Macmillan, London.

Austin, J. L. (1962), *Sense and Sensibilia*, Oxford University Press, London. V.e.: Tecnos.

Carterette, E. C. y Friedman M. P. (eds) (1974), *Handbook of Perception*, Academic Press, New York. V.e.: Trillas.

Costall, A. y Still, A. (eds). (1987), *Cognitive Psychology in question,* Harvester, Brighton.

Chisholm, R. M. (1957), *Perceiving,* Cornell University Press, Ithaca.

Churchland, P. M (1988), «Perceptual plasticity and theoretical neutrality: a reply to Jerry Fodor»: *Philosophy of Science, 55.*

Davidson, D. (1989), «The Myth of the Subjective», en M. Krausz (ed.), *Relativism,* University Press, Notre Dame. V.e.: Paidós.

Dicker, G. (1980), *Perceptual Knowledge*, Reidel Dordrecht.

Dretske, F. I. (1969), *Seeing and Knowing*, Routledge & Kegan Paul, London.

Dretske, F. I. (1981), *Knowledge and the Flow of Information*, MIT, London. V.e.: Salvat.

Fodor, J. (1983), *The Modularity of Mind*, MIT, London. V.e.: Morata.

Fodor, J. (1988), «A reply to Churchland's "Perceptual plasticity and theoretical neutrality"»: *Philosophy of Science, 55.*

García-Albea, J. E. (ed.) (1986), *Percepción y computación*, Pirámide, Madrid.

Gibson, J. J. (1966), *The Senses Considered as Perceptual Systems*, Houghton Mifflin, Boston.

Gibson, J. J. (1979), *The ecological approach to visual perception*, Houghton Mifflin, Boston.

Goldman, A. H. (1976), «Discrimination and Perceptual Knowledge»: *Journal of Philosophy, 84.*

Gregory, R. L. (1970), *The Intelligent Eye*, Weindenfeld & Nicholson, London.
Gregory, R. L. (1988), *Odd Perceptions*, Routledge & Kegan Paul, London.
Grice, H. P. (1961), «The causal theory of perception», recopilado en Warnock, 1967.
Hanson, W. H. (1969), *Perception and Discovery*, Freeman Cooper, San Francisco.
Hamlyn, D. W. (1961), *Sensation and Perception*, Routledge and Kegan Paul, London.
Hamlyn, D. W. (1990), *In and Out of the Black Box*, Basil Blackwell, Oxford.
Husserl, E. (1913), *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie*, Max Niemeyer, Halle.
Jackson, F. (1977), *Perception: A Representative Theory*, University Press, Cambridge.
Koffka, K. (1935), *Principles of Gestalt Psychology*, Harcourt, Brace & World, New York. V.e.: Paidós.
Mackie, J. L. (1976), *Problems from Locke*, University Press, Oxford.
Marr, D. (1982), *Vision*, W. H. Freeman & Company, New York. V.e.: Alianza.
Merleau-Ponty, M. (1945), *Phénoménologie de la perception*, Gallimard, Paris. V.e.: Península.
Moore, G. E. (1922), *Philosophical Studies*, Routledge & Kegan Paul, London.
Pitcher, G. (1971), *A Theory of Perception*, University Press, Princeton.
Price, H. H. (1932), *Perception*, Methuen, London.
Quine, W. O. (1975), «The Nature of Natural Knowledge», en J. Guttenplan (ed.), *Mind and Language*, Clarendon Press, Oxford.
Quinton, A. (1955), «The Problem of Perception», recopilado en Warnock, 1967.
Rivière, A. (1987), *El sujeto de la psicología cognitiva*, Alianza, Madrid.
Rock, I. (1983), *The Logic of Perception*, MIT, London.
Russell, B (1912), *The Problems of Philosophy*, Oxford University Press, London. V.e.: Labor.
Ryle, G. (1949), *The Concept of Mind*, Hutchinson, London. V.e.: Paidós.
Ryle, G. (1960), *Dilemmas*, University Press, Cambridge.
Warnock, G. J. (ed.) (1967), *The Philosophy of Perception*, University Press, Oxford. V.e.: FDC.
Segall, H. M., Campbell, D. T. y Herskovits, M. J (1966), *The Influence of Culture on Visual Perception*, Bobbs Merrill, New York.
Skinner, B. F. (1974), *About Behaviorism*, A. Knoff, New York. V.e.: Planeta Agostini.
Wittgenstein, L. (1958), *Philosophische Untersuchungen*, Basil Blackwell, London. V.e.: Crítica.
Zubiri, X. (1980), *Inteligencia sentiente*, Alianza, Madrid.

QUALIA: PROPIEDADES FENOMENOLÓGICAS

Alfonso García Suárez

I. ¿QUÉ SON LOS *QUALIA*?

1. Qualia *y cualidades secundarias*

En la filosofía de la mente contemporánea se viene empleando el término latino *qualia* (singular: *quale*) para designar las propiedades o cualidades fenomenológicas de ciertos estados y procesos mentales y, más particularmente, de nuestras experiencias y estados perceptivos. Se trata de aquellas propiedades que determinan «cómo es» tener esas experiencias o estar en esos estados. No es fácil dar ejemplos de qualia sin prejuzgar ciertas cuestiones debatidas. Como cualquier forma de introducción de tales entidades cosechará la crítica de un partido u otro, me arriesgaré a presentarlas sin demasiados escrúpulos. Imagínense que se despiertan una mañana viendo el mundo en blanco y negro, como en una vieja película. Ciertos rasgos de su experiencia visual de, pongamos por caso, un tomate maduro, habrán sufrido un cambio. Son esos rasgos introspeccionables de las experiencias —visuales o de otra índole— lo que se denomina *qualia*.

Este término es una acuñación por analogía con *quanta* (singular: *quantum*). Especificar la masa, la energía, la longitud, la extensión o el peso de una entidad por medio de una magnitud matemática es dar una respuesta a la pregunta «*¿Quantum?*». La ciencia puede ser concebida como una empresa dirigida en parte a ese cometido. Es más, desde Descartes se ha venido insistiendo en que las explicaciones científicas deben atenerse al aspecto cuantitativo de la realidad, a sus cualidades primarias.

La distinción entre *cualidades primarias*, que corresponden aproximadamente a las cantidades mensurables de las que se ocupa la ciencia,

y *cualidades secundarias* fue introducida en el s. XVII por el científico Robert Boyle y por el filósofo John Locke.

Las primeras incluyen la extensión (o tamaño), la figura (o forma), el movimiento o reposo, el número y la solidez (o impenetrabilidad) y se las supone inseparables de la materia, de modo que nuestras ideas de ellas se asemejan realmente a las cualidades correspondientes de los objetos. Las últimas incluyen el color, el sabor, el olor, el sonido y el calor o el frío y se supone que no son verdaderas cualidades de la materia, de modo que nuestras percepciones de la mismas no se asemejan a las propiedades correspondientes. La distinción puede retrotraerse a Demócrito, que afirmó que lo dulce y lo amargo, el calor y el frío y el color sólo existen por convención y que en verdad sólo existen los átomos y el vacío.

En la tradición lockeana las cualidades secundarias se definen como aquellas cualidades cuya ejemplificación en un objeto consiste en un *poder o disposición* del objeto de producir en los sujetos perceptores experiencias sensoriales. Colin McGinn (1983) señala que el punto esencial de la tesis disposicional es que el criterio determinante de si un objeto tiene un determinado color, olor, gusto, etc., es cómo parece, huele, sabe, etc., a quienes lo perciben. Así las cualidades secundarias son subjetivas en el sentido de que los hechos experienciales son *constitutivos* de su presencia. McGinn extrae varias consecuencias de este análisis disposicional de las cualidades secundarias. En primer lugar, la completa relatividad de tales cualidades: la misma cosa puede parecer de diferentes maneras a diferentes perceptores. Otra consecuencia es que el escepticismo queda excluido por principio para las cualidades secundarias. No podríamos estar equivocados acerca de las cualidades secundarias de las cosas, ya que su presencia está constituida por cómo aparecen. Además las cualidades secundarias no forman géneros naturales. Contra lo que sostienen Thomas Reid, David Armstrong y Saul Kripke, no son reducibles a «esencias reales». No es posible identificar un color, por ejemplo, con una cierta longitud de onda. Finalmente, las cualidades secundarias son en dos sentidos explicativamente inútiles. Primero, porque no se adscriben a las cosas como parte de la empresa de explicar las interacciones causales. Las cualidades que figuran en tales explicaciones son cualidades primarias. De ahí que las ciencias físicas traten sólo de éstas. Y, segundo, porque las cualidades secundarias no explican nuestra percepción de los objetos. Dado que ser rojo se analiza en términos de parecer rojo, sería circular explicar por qué algo parece rojo diciendo que es rojo.

2. *Contenido intencional* vs. *contenido cualitativo*

Las cualidades secundarias no son propiedades de la experiencia de los objetos, sino propiedades de los objetos de experiencia. Son aquellas

propiedades de los objetos consistentes en poderes de producir —en virtud de sus cualidades primarias particulares— ciertas experiencias en las mentes de los observadores. Son, en suma, propiedades disposicionales de los objetos. Los *qualia*, *tal como los hemos definido*, son por el contrario rasgos de nuestras experiencias.

Ahora bien, no son cualesquiera rasgos de nuestras experiencias. Cuando usted se apercibe de que ahí hay un tomate rojo maduro, su experiencia es *acerca de* un determinado objeto. Así su experiencia tiene rasgos *representacionales*: representa semánticamente las cosas siendo de un cierto modo. Cualquier cosa acerca de la cual una experiencia o un estado mental es su *objeto intencional* y el modo en que una experiencia o estado mental representa las cosas constituye su *contenido intencional* o *representacional*. La matización «intencional» es importante porque la experiencia de un tomate maduro puede ser verídica o no. Tal vez usted esté siendo víctima de una ilusión o de una alucinación, pero, en cualquier caso, usted es consciente de algo, de un objeto intencional, real o no.

Pues bien, los rasgos intencionales de nuestras experiencias no son qualia. Los qualia son rasgos *cualitativos* o *fenoménicos* que se supone que subyacen a esos rasgos intencionales. Mientras que el contenido intencional de una experiencia particular de un tomate rojo es un asunto del modo en que esa experiencia representa el mundo, su *contenido cualitativo* es un asunto de *cómo* es tenerla. ¿Pero hay buenas razones para trazar esa distinción o se trata, como muchos enemigos de los qualia sugieren, de una distinción sin una diferencia?

En primer lugar, hay ciertos géneros de fenómenos mentales, como las sensaciones corporales, que tienen rasgos cualitativos pero carecen de contenido representacional. Así la experiencia de un dolor comporta la percepción de ciertas cualidades intrínsecas, pero, a diferencia de una experiencia perceptiva, no representa el mundo externo. No es que los dolores carezcan de causas externas, es que la sensación dolorosa no representa semánticamente esas causas.

Una segunda razón para la distinción entre estos dos géneros de contenido nos la proporcionan los casos tipo Molyneux. En una carta a Locke, reproducida en la segunda edición del *Ensayo sobre el entendimiento humano*, William Molyneux planteó la siguiente cuestión:

> Supongamos un Hombre ciego de nacimiento y al que se le ha enseñado a distinguir por su tacto entre un Cubo y una Esfera... Supongamos luego... que al Ciego se le hace ver. Pregunto si por su vista, antes de tocarlos, podría ahora distinguir y decir cuál es el Globo y cuál el Cubo (Locke, 1690, II.ix.8).

Molyneux dio una respuesta negativa —suscrita por Locke y Berkeley— a su pregunta y vio en ello una prueba de que las ideas de forma y

contorno visible no podrían adquirirse por tacto, razonamiento o cualquier cosa que no fuera la experiencia visual. El problema de Molyneux ya no es un mero experimento mental. Los trabajos empíricos sobre recuperación de la vista realizados por R. L. Gregory y otros han aportado nuevos datos cuya relevancia para la cuestión de Molyneux es objeto de discusión. En cualquier caso, parece plausible que algunos ciegos de nacimiento captarán, al lograr la vista, diferencias y semejanzas entre experiencias visuales que durante algún tiempo no tienen significado representacional para ellos. Así Gregory (1974) informa de un sujeto que, tras recuperar la visión, no tenía impresión de profundidad ante el cubo de Necker y otros dibujos semejantes.

Tenemos, en tercer lugar, la hipótesis del espectro invertido y otros experimentos mentales sobre inversión de qualia. De nuevo hemos de recurrir a Locke, que, en el *Ensayo*, contempló la posibilidad de que

> *el mismo Objeto produjese en las Mentes de diversos Hombres diferentes* Ideas al mismo tiempo; *v.g.* si la *Idea* que produjese una *Violeta* en la Mente de un Hombre por medio de sus Ojos fuera la misma que produjese una *Caléndula* en la de otro Hombre, y *viceversa* (Locke, 1690, II.xxxii.15).

Un caso especialmente interesante es la inversión *intra* personal. Supongamos que un individuo, Fredo, sufre inversión de los colores en el tiempo *t*. El cielo le parece rojo; los tomates maduros, azules; etc. Pero Fredo reacciona al cambio y se habitúa a decir que un objeto le parece rojo en exactamente las mismas circunstancias en las que lo dirían las personas normales. Si un tomate maduro le parecía rojo antes de *t* y después de *t*, entonces el tomate le parecía del «mismo» color en el sentido de que sus experiencias tenían el mismo contenido representacional —eran *del* mismo objeto de color o tenían el mismo color como su objeto intencional—. Pero en otro sentido el tomate no le parece el «mismo» después de *t*: el contenido cualitativo de sus experiencias ha variado como resultado de su inversión del espectro.

En la hipótesis del espectro invertido contemplamos la posibilidad de una inversión de rasgos cualitativos acompañada de constancia intencional. Ned Block ha propuesto recientemente un experimento mental de efecto inverso: constancia cualitativa junto con inversión intencional. Así los casos de Tierra Invertida o Gemela por él propuestos ofrecen una razón más para la distinción que nos ocupa. Volveremos más adelante sobre estos casos.

3. *Accesibilidad a la conciencia*

Los qualia presentan *prima facie* problemas al fisicismo —la tesis de que los conceptos y las leyes de la física son en principio suficientes para dar

cuenta de todos los fenómenos naturales—. La reacción de algunos qualófobos, como Gilbert Harman (1989,1990), ha consistido en negar la accesibilidad a la conciencia de los rasgos cualitativos. Harman sostiene que si los qualia se definen, tal como hemos hecho, como rasgos de experiencias particulares más bien que como rasgos de los objetos de experiencia, entonces el problema es que no somos conscientes de ellos. Volviendo a la experiencia particular que usted tiene de un tomate maduro, Harman afirmaría que usted puede ser consciente de rasgos intencionales de ella, tales como que es la experiencia de ver un tomate rojo maduro, pero no es obviamente consciente de aquellos rasgos de su experiencia en virtud de los cuales es una experiencia con ese rasgo intencional. Harman se apresura a puntualizar que, si se objeta que el rojo del cual es usted consciente debe ser un rasgo de algo mental porque usted puede ser consciente de él incluso cuando no hay ningún tomate físico ante usted (en una alucinación, por ejemplo), entonces se comete la falacia del «argumento de la ilusión» en favor de la teoría de los datos sensoriales, *i.e.*, la falacia de postular entidades vicarias con rasgos cualitativos intrínsecos para mediar entre nuestra percepción y los objetos percibidos.

En réplica a Harman, Sidney Shoemaker (1991) sostiene que es posible mostrar que la apercepción que tenemos del contenido intencional de nuestra experiencia comporta una apercepción de qualia. Su estrategia consta de dos pasos. Muestra, en primer lugar, que aquella apercepción involucra la consciencia de un tipo de semejanza entre experiencias —semejanza cualitativa— que no puede equipararse a la semejanza intencional. Pues consideremos una situación en la que tenemos experiencia de dos objetos indistinguibles. Los objetos nos parecen iguales en el sentido de que nuestras experiencias de ellos comparten las mismas propiedades intencionales. Ambas son, pongamos por caso, «de rojo». Así, en virtud de ello están en una relación de *semejanza intencional*: son «de», o representan, el mismo tono de color. Es también verdad que son *exactamente iguales fenoménicamente*. Pero es claro que ambas representan el mismo tono de color, *i.e.*, son intencionalmente similares, *porque* son fenoménicamente iguales, algo que no podría suceder si ser fenoménicamente iguales sólo significase que representan el mismo tono de color. Tenemos, pues, un caso en el que la conciencia de la semejanza intencional entre ciertas experiencias depende constitutivamente de la conciencia de su semejanza cualitativa. Shoemaker había expresado anteriormente este punto afirmando que «la noción de semejanza de experiencia debe entenderse en términos de la noción más fundamental de experiencia *de* semejanza» (Shoemaker, 1975a, 179).

Alguien podría concederle este punto a Shoemaker y sin embargo negarse a dar el paso siguiente. Pues podría replicarle que, aunque las experiencias sean fenoménicamente semejantes o diferentes en virtud de sus

rasgos no intencionales, no hay necesidad de suponer que debamos ser conscientes de esos rasgos a fin de ser conscientes de las semejanzas y diferencias que se dan en virtud de ellos. Shoemaker, en cambio, cree que la aceptación del primer paso «requiere que las experiencias tengan rasgos no intencionales, de los que somos conscientes, en virtud de los cuales están entre sí en relaciones de semejanza y diferencia fenoménica. Estos serán los *qualia*» (Shoemaker, 1991, 516).

El qualófilo no tiene por qué cometer la falacia del dato sensorial, como Harman sugiere. Ya hemos visto que el contenido intencional de la experiencia puede no reflejar lo que hay realmente en el mundo. Cuando Macbeth sufre la alucinación de una daga, su experiencia es experiencia *de* una daga, aunque no hay tal instrumento ante él. Algunos teóricos de la percepción concluyen a partir de esta situación que, siempre que a alguien le parece ver algo cuando no hay nada ante él, se tiene que reconocer la existencia de un dato sensorial que tiene las propiedades percibidas —los *qualia* de los que él se apercibe—. Así, un análisis *acto-objeto* de la conciencia perceptiva parece comprometernos con datos sensoriales, entidades que un fisicista no puede admitir en su inventario de la realidad. Si, por el contrario, optamos por una análisis *unario* o *adverbial* de los estados perceptivos en términos de *sentires*, parece que nos libramos de los objetos internos problemáticos y sus *qualia*. Bajo tal análisis, los qualia resultan ser cualificaciones adverbiales de experiencias. U. T. Place, el iniciador del análisis adverbial, acusaba a los teóricos de los datos sensoriales de cometer la «falacia fenomenológica», la falacia de postular algo verde en mi mente cuando, por ejemplo, tengo una post-imagen verde. Si hacemos tal postulación, hemos introducido una entidad, un dato sensorial, para el que no hay lugar en el mundo físico. Pero tener una post-imagen verde, sostenía Place, es sólo tener el tipo de experiencia que tenemos cuando miramos una mancha verde. Y esas experiencias son idénticas a procesos cerebrales. No hay, por tanto, cosas llamadas «post-imágenes» que sean verdes; hay sólo experiencias-como-de-algo-verde.

Ahora bien, esta maniobra, lejos de liberar al fisicista del problema de los qualia, le abre al qualófilo otra vía para categorizarlos. Hemos visto que hay razones para pensar que ciertos estados mentales tienen carácter fenoménico intrínseco. Los qualia son rasgos de tales estados y así de la realidad. Si esos estados deben analizarse adverbialmente, entonces los qualia serán rasgos intrínsecos de los sentires y el fisicista aún estará obligado a dar cuenta de esos rasgos dentro de su concepción de lo que hay.

II. EL ARGUMENTO DE KRIPKE CONTRA LA TEORÍA DE LA IDENTIDAD PSICOFÍSICA

1. *La teoría de la identidad psicofísica*

A finales de los años 50 y en la década de los 60 la teoría de la identidad psicofísica fue la posición que gozó de mayor popularidad en filosofía de la mente. Se trata de una versión del fisicismo. A la hora de dar cuenta de los fenómenos mentales, el fisicista puede adoptar dos líneas de actuación. La línea dura es la seguida por los *materialistas eliminativos* como Paul Feyerabend, Richard Rorty, Steven Stich o Paul y Patricia Churchland. De acuerdo con esta opción, puesto que los estados mentales no pueden ser acomodados en el mundo descrito por la física, no existen. Al igual que las brujas y los fantasmas, no son más que entidades postuladas por una teoría precientífica, la psicología popular. (Se cuenta que Sir Peter Strawson comentó sarcásticamente: «Ah, sí, el dominio de gentes populares tan simples como Flaubert, Proust y Henry James».) Pero la línea más común es la ensayada por los *materialistas reductivos*: identificar los fenómenos mentales con fenómenos físicos en sentido amplio. Quienes han seguido esta segunda línea han adoptado, a su vez, uno de dos enfoques. El primero consiste en identificar los fenómenos mentales con disposiciones comportamentales. Es el enfoque *conductista* de James Watson, Rudolf Carnap o Gilbert Ryle. A finales de los 50 hubo la impresión generalizada de que el conductismo era inadecuado. Para decirlo con U. T. Place (1956), el conductismo dejaba «un residuo intratable» de conceptos mentales cuya descripción sólo parece posible en términos de procesos internos. Fue entonces cuando los materialistas reductivos ensayaron un segundo enfoque consistente en la identificación de los fenómenos mentales, no con las disposiciones a la conducta, sino con las causas de estas disposiciones. Es la vía abierta por los defensores de la *teoría de la identidad psicofísica*.

De acuerdo con ellos, hay una identidad entre *tipos* de fenómenos mentales, tales como el dolor, y *tipos* de fenómenos físicos, tales como la estimulación de las fibras C. (De ahí que se hable también de Teoría de la Identidad Tipo-Tipo.) En general, los fenómenos mentales no son disposiciones a la conducta, sino que son idénticos con estados y procesos del sistema nervioso central responsables de las disposiciones comportamentales. (De ahí el nombre «Materialismo del Estado Central».)

Fue Place (1956) el primero que sugirió que los fenómenos mentales consisten en procesos cerebrales. Place comparó esta identificación psicofísica con la identificación del relámpago con el movimiento de cargas eléctricas. El objetivo de la comparación era llamar la atención sobre el hecho de que tanto en uno como en el otro caso la identificación era el producto no de un análisis semántico o lógico de los conceptos identifi-

cados, sino de un descubrimiento empírico *a posteriori*. Y al igual que los enunciados sobre relámpagos no son sinónimos con enunciados sobre carga eléctrica, los enunciados sobre sensaciones no son sinónimos con enunciados sobre procesos cerebrales. La posición iniciada por Place fue elaborada por J. J. C. Smart y por D. M. Armstrong.

2. *El argumento de Kripke contra la teoría de la identidad tipo-tipo*

Dos influyentes trabajos de Saul Kripke (1971, 1972) sometieron a dura prueba esta teoría. Comienza Kripke señalando que las identificaciones teóricas del tipo (1) El calor es el movimiento de las moléculas, no son enunciados contingentes. Es cierto que esos enunciados expresan verdades *a posteriori*, verdades que son el fruto de un descubrimiento científico. Pero el que un enunciado valga *a posteriori* no comporta que sea un enunciado contingente. Esos enunciados, *si* son verdaderos, lo son necesariamente. La razón es que los términos singulares que en ellos aparecen son *designadores rígidos*, esto es, expresiones que designan el mismo objeto en todo mundo posible en que ese objeto existe. Y un enunciado de identidad en cuyos flancos aparezcan designadores rígidos será necesario si es verdadero, pues valdrá en todo mundo posible. Tomar (1) por contingente es confundir la distinción epistemológica *a priori/a posteriori* con la distinción metafísica necesario/contingente. En la misma condición se encuentran las identificaciones que formulan los materialistas del estado central. El enunciado (2) El dolor es la estimulación de las fibras C es un enunciado necesario, si es que es verdadero, ya que «el dolor» y «la estimulación de las fibras C» son rígidos.

Tenemos, pues, una analogía entre (1) y (2). En ambos casos, si los enunciados de identidad son verdaderos, son necesariamente verdaderos. Y en ambos casos los enunciados *parecen* contingentes porque son *a posteriori*. Pero hay una asimetría que surge cuando intentamos explicar esa ilusión de contingencia. Dado que ni (1) ni (2) son conocidos *a priori*, podemos imaginar que son falsos. ¿Cómo conciliar esa aparente contingencia con su necesidad real?

Según Kripke, esto es fácil de hacer en el caso de (1). Podemos imaginar una situación en la que el calor no hubiera producido en nosotros la sensación de calor por la que reconocemos fenoménicamente la presencia del calor. Es concebible que lo que llamamos sensación de calor pudiera haber sido producido por otra propiedad de los objetos. En tal caso estaríamos, cualitativamente hablando, en la misma situación epistémica que aquélla en la que estábamos antes de descubrir la identidad del calor con el movimiento molecular medio. Tanto en la situación imaginaria como en nuestra situación actual podemos designar el calor por medio de la descripción «la propiedad de los objetos que produce esta sensación en nosotros». Esta descripción sirve para fijar la referencia

del designador rígido «calor», pero la descripción en cuestión no es rígida. De ahí que sea contingentemente verdadero que el calor, *designado mediante esta descripción asociada*, es el movimiento de las moléculas. Esta verdad es, en expresión de Michael Levin (1975) el *descubrimiento contingente asociado* que sirve para explicar la ilusión de contingencia que produce (1).

Para explicar la aparente contingencia de (2) el materialista debe hallar un correspondiente descubrimiento contingente asociado. Debe describir una situación epistémica, indistinguible de nuestra situación efectiva antes de descubrir (2), en la que algo cualitativamente idéntico al dolor sea designado por un designador no rígido y luego ese algo resulte no ser la estimulación de las fibras C. Si esto fuera posible, sería contingentemente verdadero que el dolor, *así designado*, es idéntico a la excitación de las fibras C. Pero, sostiene Kripke, una situación epistémica que fuese cualitativamente idéntica a una situación en la que se tiene la sensación de dolor sería ella misma una situación en la que hay dolor. Imaginar una situación en la que se está experimentando una sensación cualitativamente idéntica a un dolor es imaginar una situación en la que se está experimentando dolor. A diferencia de los objetos y fenómenos físicos, para las sensaciones no hay distinción entre cómo parecen y cómo son realmente. Aquí *esse* es *percipi*. Así el materialista no puede explicar por qué (2) parece contingente y por ello su identificación resulta sospechosa.

3. *Algunas réplicas a Kripke*

El argumento de Kripke parte del supuesto de que «dolor» es un designador rígido. Ahora bien, de acuerdo con las teorías causales y funcionales el dolor puede definirse como la sensación *ocupante de un rol* determinado causal y/o funcional: el dolor es cualquier cosa que juegue un cierto papel causal y/o funcional en nuestra vida. Diferentes tipos de cosas podrían desempeñar ese rol. Si se admite esto, se sigue que «dolor» no es rígido. Lo que es de hecho dolor podría no haber sido dolor. En otro mundo posible, el dolor podría no haber sido la estimulación de las fibras C.

Otra escapatoria posible a la objeción de Kripke es la propugnada por Michael Levin (1975). Consiste en conceder que tanto «dolor» como «estimulación de las fibras C» son rígidos, pero sostener que el único modo de fijar su referencia es usando «descripciones australianas» del tipo de las propuestas por Smart, David Lewis y Armstrong. Así «dolor» designaría el fenómeno que designan no rígidamente las descripciones australianas adecuadas del tipo «la sensación causada por tales y cuales estímulos». Tenemos entonces un descubrimiento contingente asociado a la verdad necesaria (2): el descubrimiento de que los fenómenos que

responden a ciertas descripciones australianas son de hecho cierto tipo de fenómenos cerebrales.

Una salida más radical es la practicada por William Lycan (1974, 1987). Recordemos que Kripke se apoya en el supuesto de que no hay diferencia entre realidad y apariencia en el caso de una sensación como el dolor. Lycan niega esta equiparación del dolor con la impresión o apercepción de sentir dolor. Si hay argumentos plausibles contra la tesis de que el estado cognitivo que es la sensación de dolor es el dolor mismo, entonces es plausible sostener que el estado cognitivo puede darse sin que se dé el dolor. Si una persona pudiera tener la falsa impresión de tener un dolor, estaría en un estado epistémicamente indistinguible del de alguien que siente realmente un dolor. Y esto le proporcionaría el requerido descubrimiento contingentemente asociado al materialista que defiende que el dolor es necesariamente la estimulación de las fibras C. Sería el descubrimiento de que lo que usualmente produce la impresión de sentir dolor es la estimulación de las fibras C. El inconveniente de esta escapatoria es que resulta difícil de tragar la idea de que alguien puede equivocarse acerca de si él realmente siente dolor (Wittgenstein, 1953; cf. García Suárez, 1976).

III. FUNCIONALISMO Y *QUALIA*

1. *El aspecto conductista del funcionalismo*

Los funcionalistas no se comprometen con identidades tipo-tipo como las ejemplificadas por (2). Identifican, por el contrario, los fenómenos mentales con estados funcionales de orden superior que pueden ser realizados por una variedad de sistemas físicos. Para el funcionalista los estados mentales son definibles en términos de sus relaciones con *inputs* sensoriales, *outputs* comportamentales y otros estados mentales. Un dolor, por ejemplo, sería una sensación típicamente causada por un daño corporal que tiende a provocar en el que lo padece ciertas pautas de conducta —quejarse, gemir, etc.— y que suele suscitar el deseo de librarse de él.

Pero parece lógicamente posible que haya creaturas que tengan dolores y carezcan de la organización funcional requerida o que estén en el estado funcional apropiado pero no sientan dolor. Y si esto es así, ninguna teoría funcional, causal o puramente relacional de los fenómenos mentales puede dar cuenta de sus rasgos fenomenológicos introspeccionables. Dos organismos podrían compartir la misma organización funcional y sin embargo tener, por ejemplo, espectros cromáticos invertidos el uno por relación al otro —el problema de los *qualia* invertidos—. Es más, podría suceder que dos organismos fuesen funcionalmente idénticos y sin embargo uno de ellos careciera de estados mentales con contenido

cualitativo —el problema de los qualia ausentes (Sanfélix Vidarte, 1991)—. Las objeciones basadas en la posibilidad de qualia invertidos o qualia ausentes explotan el aspecto conductista del funcionalismo. Si la equivalencia funcional es equivalencia al nivel de *inputs* y *outputs* especificados no mentalmente, entonces creaturas con *qualia* invertidos o creaturas sin qualia serán equivalentes a nosotros en la medida en que lo sean en términos de tales *inputs* y *outputs*.

2. Qualia *invertidos*

La hipótesis del espectro invertido que Locke había planteado fue discutida desde comienzos de nuestro siglo por autores como C. I. Lewis, M. Schlick, H. Reichenbach, J. Wisdom, M. Black y J. J. C. Smart. Los positivistas lógicos la rechazaron aplicándole el principio de verificación. Dado que no podríamos verificar en principio si otra persona sufre o no una inversión del espectro de los colores, el supuesto de que tal inversión podría suceder sería carente de significado empírico. Pero esta disolución del problema no contentaba a quienes no admitían el principio de verificación como criterio de significatividad. Y, lo que es más importante, no sirve para el caso de la inversión intrapersonal, como veremos más abajo.

En «What Psychological States Are Not», Ned Block y Jerry Fodor (1972) resucitaron la hipótesis y la propusieron como un argumento contra el funcionalismo. Comenzaremos exponiendo la versión *inter*personal clásica del argumento. ¿Cómo sé que ustedes y yo tenemos la misma sensación visual cuando miramos el cielo despejado? Usted dice que ve azul el cielo y yo concuerdo con usted en cuanto al carácter de mi experiencia. ¿Pero cómo sé que lo que usted llama «azul» no es lo que yo llamo «rojo» y viceversa? Supongamos que nuestros respectivos espectros cromáticos estuvieran invertidos. Usted vería el cielo del color que yo veo la sangre y vería la sangre del color que yo veo el cielo; y así para todo par de colores complementarios. Puesto que ambos hemos aprendido nuestras palabras designativas de colores al sernos mostrados objetos públicos de color, nuestra conducta verbal encajaría aun cuando nuestras experiencias de los colores fuesen enteramente diferentes. Así parece que las nociones de semejanza, diferencia e identidad cualitativa carecen de sentido cuando se aplican intersubjetivamente. No es posible comparar qualia interpersonalmente. Y esto parece grano para el molino verificacionista: la idea misma de una inversión de *qualia* carece de sentido. ¿Significa esto que los qualia son reales pero incomparables o significa que no tiene sentido hablar de *qualia*? La mayoría de los filósofos se inclinaron por la primera alternativa: los *qualia* existen pero no hay modo de detectar diferencias interpersonales con respecto a ellos.

Cuando se formuló en los años 60 la versión *intra*subjetiva del problema, las cosas cambiaron. Pues supongamos que un buen día —tal vez

como resultado de algún cambio neurológico— me levanto y para mi sorpresa encuentro que el cielo me parece rojo, veo azul la sangre, etc. Ahora las experiencias que se comparan son las de (dos fases de) un solo individuo, no las de dos individuos distintos. Y aquí ya no son aplicables los escrúpulos verificacionistas, pues un cambio así sería detectable por medio de la introspección, o de la introspección más la memoria, de la persona que lo sufre. Y si la víctima del cambio puede detectarlo, otras personas podrían tener conocimiento de él a través de los informes veraces del sujeto del cambio. Es más, habría conducta no verbal, y no sólo conducta verbal, que podría indicar ese cambio. Como apunta Sidney Shoemaker:

> Afirmar que la inversión del espectro es posible pero que es indetectable incluso en el caso intrasubjetivo, sería cortar la conexión que suponemos que se da entre los estados cualitativos y la conciencia introspectiva de los mismos.y también sus conexiones con las creencias perceptivas acerca del mundo y, *via* esas creencias, con la conducta (Shoemaker, 1975b, 259).

Ahora bien, hay un argumento que lleva de la premisa de que es posible la inversión del espectro intrapersonal a la conclusión de que debe ser también posible una inversión interpersonal. Supongamos que un individuo A sufre inversión intrasubjetiva en el tiempo *t*. Bajo el supuesto de que los demás no sufren también inversión en *t*, parece que o bien antes de *t* o después de *t* (o en ambos tiempos) la experiencia de los colores de A debe ser distinta de la experiencia de los colores de los demás individuos. Pero si admitimos que puede haber tal inversión intersubjetiva en los casos en que hay inversión intrasubjetiva, tenemos que admitir también que podría haber inversión intersubjetiva sin inversión intrasubjetiva. Si es posible que la experiencia de lo colores de A sea diferente de la de sus congéneres en un tiempo determinado, debe ser posible que se dé esa diferencia *todo* el tiempo.

La hipótesis del espectro invertido es un caso particular del problema de la inversión de los qualia. La razón por la que los filósofos se han centrado en la experiencia de los colores es que en este caso el supuesto de una posible inversión es más plausible que en el caso de los *qualia* de otras experiencias. No obstante, podría suponerse que lo que yo experimento cuando afirmo sinceramente que siento dolor es diferente de lo que usted experimenta bajo las mismas circunstancias. Si el funcionalismo es correcto, las relaciones causales o funcionales de un estado mental con *inputs* sensoriales, *outputs* conductuales y otros estados mentales habrían de mantenerse constantes en su caso y en el mío. Pero esto es compatible con la posibilidad de que el *quale* de lo que yo llamo «dolor» sea enteramente distinto del quale de lo que usted llama así. En tal caso el funcionalismo quedaría en entredicho. Ninguna especificación de la or-

ganización funcional de un organismo podría bastar para determinar los rasgos cualitativos del dolor o de cualquiera otra experiencia.

Block y Fodor no consideran que la posibilidad de una inversión de los qualia sea una objeción decisiva al funcionalismo, pues creen que el funcionalista podría negar el supuesto de que los dolores, pongamos por caso, deban ser cualitativamente semejantes. El funcionalista podría afirmar que «el carácter de los *qualia* de un organismo es irrelevante para si siente dolor o (equivalentemente) que los dolores son sentidos de modo diferente por diferentes organismos» (Block y Fodor, 1972, 245). Esto es tanto como renunciar al funcionalismo como teoría del *aspecto cualitativo* de la experiencia, aunque retenerlo como teoría del *aspecto cognitivo* de la mente. El problema de esta solución es que abre la puertas a la posibilidad de *qualia* ausentes. Pues si se admite que un estado funcional dado podría existir sin tener un contenido cualitativo *particular*, ¿no deberíamos admitir también que pudiera existir sin tener *ningún* contenido cualitativo en absoluto?

3. Qualia *ausentes*

Block y Fodor han propuesto otro *Gedankenexperiment* en el que se contempla la posibilidad de un organismo funcionalmente idéntico a nosotros pero carente de rasgos cualitativos asociados a sus estados funcionales. En«Troubles with Functionalism» Block (1978) se imagina un cuerpo que es externamente como el cuerpo humano pero cuyo interior es enteramente diferente. En su cerebro hay un ejército de homúnculos, cada uno de los cuales tiene una tarea específica: implementar un cuadrado de una máquina de Turing que describe a ese individuo. Por medio de los esfuerzos combinados de los homúnculos, el sistema realiza la misma tabla de máquina que un individuo y así es funcionalmente idéntico a él. Por si alguien objetara que se necesitarían muchísimos hombrecillos para esa tarea colectiva, Block propone un segundo caso en el que los homúnculos son reemplazados por la nación china:

> Supongamos que convertimos al gobierno de China al funcionalismo y convencemos a sus autoridades de que realzaría enormemente su prestigio internacional realizar una mente humana por una hora. Le proporcionamos a cada una del millar de millones de personas de China (elegí China porque tiene un millar de millones de habitantes) un radiotransmisor especialmente diseñado que la conecta del modo apropiado con otras personas y con el cuerpo artificial mencionado en el ejemplo previo. Reemplazamos los hombrecillos por un transmisor y receptor conectados con las neuronas de entrada y salida [...] disponemos letras exhibidas en una serie de satélites colocados de manera tal que puedan ser vistos desde cualquier lugar de China. El sistema de un billón de personas comunicándose entre sí más los satélites desempeña el papel de un «cerebro» externo conectado por radio con el cuerpo artificial [...] No es en absoluto obvio que el sistema China-cuerpo sea físicamente im-

posible. Podría ser funcionalmente equivalente a usted por un breve tiempo, pongamos una hora (Block, 1978, 276).

J. R. Searle (1980, cf. 1983, 1984, 1992) ha propuesto otra fantasía del mismo tipo. Imaginemos un solo homúnculo que está dentro de una habitación con aberturas de *input* y *output*. Se le ha proporcionado un manual que codifica el programa de un ser consciente, concretamente de un hablante nativo de chino. Al recibir *inputs*, el homúnculo, a gran velocidad, consulta su manual y con lápiz y papel calcula el *output* adecuado. Nadie sugeriría que el homúnculo entiende chino, como nadie sugeriría que el cuerpo artificial de Block tiene experiencias conscientes. Sin embargo, el programa es «realizado» en el sentido funcionalista.

4. *Compatibilismo e incompatibilismo*

Ante los problemas planteados por la posibilidad de qualia invertidos y de qualia ausentes caben tres posiciones que, siguiendo a Shoemaker, podemos llamar incompatibilismo funcionalista, incompatibilismo anti-funcionalista y compatibilismo, respectivamente. El incompatibilista funcionalista salva el funcionalismo negando la existencia de *qualia*. El incompatibilista anti-funcionalista admite la existencia de *qualia* y deduce de ello que el funcionalismo es falso o, al menos, incompleto. El compatibilista trata de conciliar la admisión de los *qualia* con el funcionalismo.

a) Incompatibilismo funcionalista: Dennett sobre cómo «quinear» los *qualia*.

El incompatibilismo funcionalista es la posición más radical. Sus principales exponentes son Gilbert Harman y Daniel Dennett. Puesto que ya hemos atendido a la posición del primero en 1.3, nos centraremos en el último. Dennett (1988, 1991) propone «quinear» los *qualia*. «Quinear» es un verbo humorístico acuñado en honor de W. V. Quine que significa negar resueltamente la existencia o importancia de algo aparentemente real o significativo. Dennett niega la existencia de qualia aunque admite que *parece* haberlos. Parece haberlos porque la ciencia ha mostrado que las cualidades secundarias, los colores por ejemplo, no son propiedades objetivas que estén ahí fuera y sentimos por tanto que deben estar aquí dentro de nuestras mentes. Pero, arguye Dennett, lo que la ciencia muestra es sólo que las propiedades reflectantes de la luz de los objetos causan que las creaturas estén en diversos estados discriminativos. Estos estados discriminativos de los cerebros de los observadores tienen varias cualidades primarias. Y en virtud de ellas tienen varias propiedades secundarias meramente disposicionales. ¿No tienen también nuestros estados discriminativos algunas propiedades «intrínsecas», subjetivas, privadas e

inefables que constituyen el modo en que las cosas nos aparecen? No, responde Dennett: «Cuando dices *Este* es mi quale, lo que estás identificando, o refiriendo, *tanto si lo adviertes como si no*, es tu complejo idiosincrático de disposiciones» (Dennett, 1991, 388).

Cabría replicar que mis disposiciones podrían cambiar sin que cambiase mi quale intrínseco o que mis qualia podrían invertirse sin inversión de todas mis disposiciones. Tal es la hipótesis del espectro invertido. Dennett señala que, para que la hipótesis funcionara como contraejempolo al funcionalismo, el qualófilo tendría que describir un experimento mental que demostrase que la apariencia de las cosas puede ser independiente de todas las disposiciones reactivas al color. Sería necesario imaginar un caso en el que se diera la inversión de *qualia y a la vez* algo anulara un cambio de disposiciones reactivas del sujeto. En la literatura sobre el espectro invertido se supone que la segunda conmutación se realiza por adaptación gradual. Suponga que mientras duerme sufre usted una intervención quirúrgica que invierte su visión de los colores y que *todas* sus disposiciones reactivas se restauran gradualmente. ¿Cuál sería su intuición acerca de sus qualia? ¿Están aún invertidos o no? Quizá le parezca que sus qualia siguen invertidos. Dennett sugiere entonces que la explicación más probable de esta intuición es que «está usted haciendo el supuesto adicional, no garantizado, de que toda la adaptación está sucediendo del "lado post-experiencial"» (Dennett, 1991, 394). ¿Pero no podría ocurrir en el lado pre-experiencial? Supongamos que completamos el experimento mental añadiendo que, a medida que se produce la adaptación, usted encuentra que los colores de las cosas ya no le parecen extraños y a veces se encuentra confuso y hace dobles correcciones en sus disposiciones verbales. Cuando se le pregunta por el color de un objeto dice «Es verd... —no rojo— no ¡es verde!» Ahora podría parecer obvio que los qualia de color mismos se han adaptado o se han reinvertido. Pero, tendemos a pensar, tendrá que ser lo uno o lo otro. Esa convicción, afirma Dennett, se basa en el supuesto no examinado de que todas las adaptaciones pueden categorizarse como pre-experienciales o post-experienciales. Pero ese supuesto sólo vale para casos extremos.

Para mostrarlo, Dennett nos enfrenta a casos como el siguiente. Mucha gente reconoce que el gusto por la cerveza es un gusto adquirido. Uno se adiestra gradualmente para disfrutar del sabor. ¿Qué sabor? ¿El del primer trago? Aquí aparecen dos bebedores de cerveza. El primero dice: «Si mi primer trago de cerveza me hubiera sabido como me sabe mi trago más reciente, nunca hubiera tenido que adquirir inicialmente el gusto por la cerveza». El otro bebedor dice: «La cerveza me sabe ahora como me supo siempre, sólo que ahora me gusta más *ese mismo sabor*». Hay aquí una diferencia en heterofenomenología que necesita explicación. Dennett afirma:

> Tenemos que mirar, más allá de los mundos hetero-fenomenológicos, a los eventos que suceden en la cabeza para ver si hay una interpretación preservadora de la verdad (aunque «forzada») de las afirmaciones de los bebedores de cerveza, y *si* la hay será sólo porque decidamos *reducir* «el cómo sabe» a un complejo de disposiciones reactivas. Tendríamos que «destruir» los *qualia* a fin de «salvarlos»(*o. c.*, 396).

Y aplicando la idea a la inversión de *qualia*:

> La idea de que hay algo *además* de la inversión de todas nuestras disposiciones reactivas, de modo que si éstas se normalizaran quedarían los qualia invertidos, es simplemente parte del tenaz mito del Teatro Cartesiano... *Si* no hay *qualia* por encima de la suma total de disposiciones a reaccionar la idea de mantener los *qualia* constantes, a la vez que se ajustan las disposiciones, es contradictoria (*o. c.*, 398).

Dennett no niega que hay «modos en que las cosas nos aparecen». Su blanco de ataque es una cierta concepción filosófica de los *qualia* según la cual éstos tienen propiedades problemáticas —son intrínsecos, atómicos, inanalizables, no relacionales, inefables, privados e inmediata e incorregiblemente aprehensibles—. Owen Flanagan (1992) reacciona afirmando que, bajo esta concepción problemática, los qualia que Dennett quinea merecen ese tratamiento. Pero, bajo la concepción ortodoxa, «qualia» nombra lo que Dennett dice que nombra —los modos en que nos aparecen las cosas— y esto no entraña ninguna de las propiedades problemáticas. Los modos en que nos aparecen las cosas nos proporcionan los tipos de explicación subjetiva tal como son discernibles desde el punto de vista de la primera persona. Y esta tipología es condición previa para la construcción teórica y la generación de hipótesis sobre la experiencia subjetiva. Volvamos a las hipótesis en competencia de que la diferencia de gustos entre los dos bebedores de cerveza se deben, o bien a un cambio en la experiencia gustativa producida por la cerveza a la vez que se mantienen las preferencias gustativas, o a un cambio de preferencias manteniéndose constante la experiencia gustativa. Dennett cree que estas hipótesis pueden elaborarse de modo tal que no hay manera de decidir entre ellas sobre la base de la mera introspección. Flanagan, en cambio, cree que es posible obtener evidencia en favor de una u otra hipótesis de fuentes distintas de la introspección. Para él «la fenomenología, la psicología y la ciencia cerebral son *partenaires* creíbles en el esfuerzo por penetrar los *qualia* y revelar la naturaleza, la estructura y los roles causales de los diversos modos en que las cosas aparecen» (Flanagan, 1992, 85).

b) Incompatibilismo antifuncionalista: el cuasi-realismo cuasi-funcional de Block

Los representantes de esta posición mantienen que los *qualia* son funcionalmente indefinibles y que esta indefinibilidad supone una grave

amenaza o al menos una restricción para el funcionalismo. Block y Fodor, por ejemplo, sostienen que, si las situaciones contempladas en los experimentos mentales antes relatados son concebibles, ello mostraría que algo que es crítico para ciertos estados mentales —el «cómo es ser» un organismo que tiene experiencias— no sería captado en las relaciones causales que figuran en los análisis funcionalistas.

Recientemente Ned Block (1990) ha descrito su posición como cuasi-realismo cuasi-funcional. Es cuasi-*realista* sobre qualia porque se compromete con la existencia de rasgos mentales intrínsecos de nuestra experiencia y en este sentido se opone al escepticismo sobre qualia de Dennett y Harman. Y es *cuasi*-realista porque no comete la falacia de intencionalizar los qualia, esto es, de suponer que los contenidos experienciales expresables en un lenguaje público son contenidos cualitativos. Pero su postura es *cuasi*-funcionalista porque Block considera que mientras que el contenido *intencional* de la experiencia es funcional, el contenido *cualitativo* no es caracterizable funcionalmente. Dos experiencias pueden diferir funcionalmente —tener diferentes contenidos intencionales— pero tener el mismo contenido cualitativo, y dos experiencias pueden ser funcionalmente idénticas —tener el mismo contenido intencional— pero tener diferentes contenidos cualitativos.

Block propone un nuevo argumento de inversión en favor del realismo sobre los *qualia* y en contra del funcionalismo estricto: el argumento de la Tierra Invertida. Es un caso de inversión intencional y funcional con constancia de contenido cualitativo. Es, pues, el caso inverso del contemplado en la hipótesis del espectro invertido, en donde teníamos un ejemplo de inversión cualitativa acompañada de identidad funcional e intencional.

La Tierra Invertida sólo difiere de la Tierra en dos aspectos: *(i)* allí todo tiene realmente el color complementario del color terrestre y *(ii)* el vocabulario de los habitantes de ese planeta también está invertido. Las dos diferencias se cancelan mutuamente en el sentido de que el discurso sobre colores en la Tierra Invertida sonaría igual que el discurso sobre colores en la Tierra. Block presenta entonces un análogo intrapersonal y un análogo interpersonal de la hipótesis del espectro invertido. Por brevedad atenderemos sólo al caso intrapersonal, que es siempre el más convincente.

Un equipo de científicos lo duerme a usted, le coloca lentes inversoras de los colores, cambia el color de su piel, etc. Sin que lo advierta es usted transportado a la Tierra Invertida y una contrapartida suya es substituida por usted. Se despierta y, puesto que las lentes inversoras cancelan los colores invertidos, no advierte la diferencia. El contenido cualitativo de su experiencia es el mismo que el día antes. ¿Qué pasa con el contenido intencional? La raíz causal de sus términos para los colores está en la Tierra. Así, en su primer día en la Tierra Invertida, sus contenidos inten-

cionales siguen siendo los mismos. Pero a medida que pasa el tiempo llegará a dominar su incorporación al entorno físico y lingüístico de la Tierra Invertida y sus contenidos intencionales cambiarán hasta ser idénticos con los de los nativos. Y lo mismo sucederá con sus estados funcionales. Pasados cincuenta años, usted y su fase anterior ejemplificarían un caso de inversión funcional e intencional acompañada de los mismos contenidos cualitativos. Y esto, concluye Block, refuta la teoría funcionalista del contenido cualitativo y establece firmemente la distinción contenido intencional/contenido cualitativo. Hay una ventaja adicional sobre el caso intrasubjetivo del espectro invertido. En este caso, el cambio *interno* del sujeto vuelve vulnerables a la duda sus informes en primera persona. Pero el sujeto del caso de la Tierra Invertida no sufre ningún cambio interno. Toda la inversión tiene lugar *fuera* del cerebro, en su entorno.

c) Compatibilismo funcionalista

A un funcionalista de pura raza que admita la existencia de qualia le están abiertas al menos dos salidas: o bien intentar dar cuenta de los qualia funcionalmente o mantener que los qualia se deben en parte a la substancia física que realiza las funciones mentales y, por ello, no son algo que el funcionalismo esté obligado a explicar.

c. 1) El compatibilismo de Shoemaker

La primera salida ha sido practicada por Sidney Shoemaker (1975a, 1975b, 1980, 1981, 1991). Comienza admitiendo que si los qualia ausentes fueran posibles, el carácter cualitativo de la experiencia sería inaccesible a la introspección. Ahora bien, cualquier estado funcionalmente idéntico a un dolor tenderá a producir en la persona que está en él, además de una cierta conducta y una creencia en que hay algo que va mal en su organismo, ciertas *creencias cualitativas*, i.e., tenderá a hacerle pensar que siente un dolor que tiene un cierto carácter cualitativo desagradable. Si los qualia ausentes fueran posibles, tal estado podría carecer de carácter cualitativo. Pero esto no es plausible:

> Puesto que uno de los rasgos causales del dolor es que es accesible a la introspección —esto es, que suscita bajo ciertas circunstancias una creencia en su ocurrencia por parte de su poseedor—, tendría que ser verdadero del dolor *ersatz* que suscita las mismas creencias introspectivas por parte de su poseedor, así como la misma conducta, que suscita el dolor. El quid de mi argumento contra la posibilidad de qualia ausentes era que la suposición de que es posible el dolor *ersatz* suscita insuperables dificultades epistemológicas. Hay, ciertamente, el problema de cómo podríamos saber que nuestros amigos y vecinos no son hombres de imitación. Por de pronto, parece que mis fundamentos para creer que mis propios dolores son reales

y no *ersatz* no pueden ser mejores que los fundamentos que un hombre de imitación tendría para creer lo mismo sobre sus dolores *ersatz* (Shoemaker, 1980, 315-316).

A esto el escéptico sobre qualia podría objetar que no hay rasgos cualitativos de la experiencia que sean accesibles a la conciencia. Ya hemos visto más arriba que Gilbert Harman defiende la tesis de que las únicas propiedades de las experiencias a las que tenemos acceso son las propiedades intencionales. Y hemos examinado también la réplica de Shoemaker, para quien la conciencia que tenemos del contenido intencional de nuestras experiencias comporta una conciencia de un tipo de semejanza entre experiencias —semejanza fenoménica— que no puede identificarse con la semejanza intencional. ¿Pero no podemos parafrasear el discurso sobre semejanza fenoménica de experiencias en términos del discurso sobre cosas que parecen o saben igual? Shoemaker reconoce que el vocabulario en términos de «parece», «sabe», etc., está hecho para describir el contenido intencional de las experiencias. Así, a fin de referirnos al contenido cualitativo, necesitamos los conceptos de quale y de semejanza fenoménica, los cuales *son conceptos teóricos*, no conceptos de sentido común. Pero son necesarios para poner de manifiesto algo que está implícito en la psicología popular. Los necesitamos para dar sentido a la batería de conceptos que la gente emplea efectivamente en sus juicios introspectivos; su aplicabilidad es una condición necesaria de la aplicabilidad de conceptos, como el de «parecer lo mismo», que la gente emplea efectivamente (Shoemaker, 1991, 521).

Shoemaker piensa en cambio que la concebibilidad de la inversión del espectro intrasubjetivo nos obliga a concluir que no podemos definir funcionalmente estados cualitativos *particulares*. Sin embargo, esto no amenaza en su opinión al funcionalismo, dado que es posible ofrecer una caracterización de la *clase* de los estados cualitativos: podemos definir funcionalmente las condiciones de identidad, semejanza y diferencia cualitativas entre los miembros de esa clase. Ahora bien, al suponer que la inversión intrasubjetiva del espectro es *detectable*,

estamos suponiendo algo... sobre las relaciones de semejanza y diferencia cualitativas, a saber, que cuando se dan entre experiencias co-conscientes, ello tiende a dar lugar a una consciencia introspectiva de la incidencia de esas relaciones mismas, i.e., tiende a dar lugar a «creencias cualitativas» correctas al efecto de que se dan esas relaciones (Shoemaker, 1975b, 199).

Y añade:

Mi sugerencia es que lo que hace de una relación entre experiencias la relación de semejanza cualitativa (fenomenológica) es precisamente el que desempeñe un cierto rol «funcional» en la conciencia perceptiva de semejanzas objetivas, a saber, que tienda a producir creencias perceptivas al efecto de que se dan tales semejanzas (*o. c.*, 199-200).

c. 2) Los *qualia* como ocupantes de roles

La segunda estrategia consiste en dar cuenta del aspecto cualitativo apelando a las estructuras físicas en las que se realizan los estados funcionales. De este modo, los *qualia* dejarían de ser algo de lo que deba dar cuenta el funcionalismo.

Paul Churchland (1984) reconoce que las realizaciones físicas de los estados funcionales tienen verdaderamente una naturaleza intrínseca de la que depende nuestra identificación introspectiva de esos estados, pero sostiene que esa naturaleza intrínseca no es esencial para la identidad-tipo de un determinado estado mental y de hecho puede variar entre un caso y otro del mismo estado mental. En la medida en que los estados mentales cualitativos son provocados por objetos rojos y son la causa de que creamos que algo es rojo, esos estados son sensaciones de rojo sea cual sea su carácter intrínseco. Los rasgos cualitativos intrínsecos tan sólo facilitan la rápida identificación introspectiva de las sensaciones. Pero no son esenciales para determinar la identidad de los estados mentales en cuanto tipos. El materialista puede hacer lugar para los *qualia* en su esquema de la realidad identificándolos con propiedades físicas de cualquier tipo de estado físico que ejemplifique los estados mentales (funcionales) que los exhiben.

David Lewis (1980) da un paso más allá y defiende que el carácter intrínseco de la experiencia es el carácter intrínseco del ocupante de un rol funcional. Pero si en respuesta a las objeciones al funcionalismo basadas en la posibilidad de *qualia* invertidos y *qualia* ausentes, identificamos los *qualia*, no con estados funcionales, sino con estados neurofisiológicos que desempeñan los roles funcionales apropiados, parece que nos vemos retrotraídos a la posición del materialismo del estado central, en cuyo caso estamos necesitados de algún recurso para desactivar el argumento de Kripke.

IV. EL RESTO ES SUBJETIVIDAD

Los argumentos que hemos examinado hasta aquí iban dirigidos contra versiones particulares del fisicismo. El argumento de Kripke estaba diseñado para desacreditar la teoría de la identidad psicofísica. Los argumentos de los qualia invertidos y los qualia ausentes eran propuestos como objeciones al funcionalismo. Examinaremos ahora algunos argumentos «de amplio espectro» que pretenden presentar un desafío al materialismo en cualquiera de sus versiones.

1. *El argumento del conocimiento*

Si las condiciones de identidad y semejanza entre qualia fueran definibles funcionalmente, como Shoemaker cree, sería legítimo hablar de ellos en cuanto realizados físicamente en el sentido en el que una propiedad funcional puede ser realizada físicamente. En tal caso, podría identificarse un quale con la disyunción de sus posibles realizaciones. La principal objeción a esta identificación es el llamado argumento del conocimiento, de acuerdo con el cual conocer todos los hechos físicos acerca de la visión de los colores, pongamos por caso, no sería suficiente para saber cómo es ver algo rojo. Así es como Frank Jackson presenta el argumento:

> Mary es una brillante científica que está... obligada a investigar el mundo desde una habitación blanca y negra a través de un monitor de televisión en blanco y negro. Se especializa en la neurofisiología de la visión y adquiere, supongamos, toda la información física que hay que obtener acerca de lo que sucede cuando vemos tomates maduros, o el cielo, y usamos términos como «rojo», «azul», etc... ¿Qué sucederá cuando Mary sea liberada de su habitación blanca y negra o se le dé un monitor de televisión en color? ¿*Aprenderá* algo o no? Parece bien obvio que aprenderá algo sobre el mundo y nuestra experiencia visual; entonces es ineludible que su conocimiento previo era incompleto. Pero ella tenía *toda* la información física. *Ergo* hay más que tener que eso y el fisicismo es falso (Jackson, 1982, 128).

Si el argumento del conocimiento fuese válido, refutaría el mínimo común denominador de todas las teorías materialistas. El mínimo neto del materialismo es una tesis de superveniencia: no hay diferencia real sin diferencia física. Pero si hubiera algún tipo de información —información fenomenológica— que puede eliminar posibilidades que son dejadas abiertas por cualquier cantidad de información física, entonces debería haber posibilidades de que sean físicamente iguales pero no sean estrictamente iguales. Así, si hubiera información fenomenológica no reducible a información física, la tesis mínima de superviniencia quedaría refutada.

En su réplica a los críticos, Jackson (1986) ofrece un modo conveniente de exponer el argumento:

> P1 Mary (antes de su liberación) conoce todo lo físico que hay que conocer acerca de otras personas.
> P2 Mary (antes de su liberación) no conoce todo lo que hay que conocer acerca de otras personas, porque aprende algo sobre ellas al ser liberada.
> C Por tanto, hay verdades sobre otras personas (y sobre ella misma) que se escapan a la descripción fisicista.

Jackson nos presenta a una Mary conocedora de todos los hechos físicos sobre nosotros y nuestro entorno

> en un amplio sentido de «físicos» que incluye todo lo que hay en una física, química y neurofisiología *completadas*, y todo lo que hay que saber acerca de los hechos causales y relacionales que se siguen de todo esto, incluyendo ciertamente los roles funcionales. Si el fisicismo es verdadero, ella conoce todo lo que hay que conocer... Los fisicistas deben mantener que el conocimiento físico completo es conocimiento completo *simpliciter* (Jackson, 1986, 392).

El *quid* de su argumento es éste:

> El problema para el fisicismo es que, después de que Mary vea su primer tomate maduro, se dará cuenta de cuán empobrecida había sido *todo el tiempo* su concepción de la vida mental de los *demás*... Pero ella conocía todos los hechos físicos sobre ellos todo el tiempo; por tanto, lo que no conocía hasta su liberación no es un hecho físico acerca de sus experiencias. Pero es un hecho acerca de ellas. Este es el problema para el fisicismo... Hay verdades acerca de otras personas (y de sí misma) que se escapan a la descripción fisicista (*o. c*, 393).

2. *Algunas réplicas al argumento de Jackson*

a) ¿Adquiere Mary algún tipo de conocimiento?

Pasaremos ahora revista a algunas respuestas al problema planteado por Jackson. La cuestión central es: ¿adquiere Mary algún tipo de conocimiento cuando experimenta la sensación de rojo por primera vez? Tan sólo Paul Churchland (1985, 1990) y Daniel Dennett (1991) se han resistido a responder afirmativamente a esta pregunta. La posición de Churchland es, sin embargo, ambigua, pues, si bien niega que Mary llegue a conocer nuevos hechos, admite que conoce de un nuevo modo hechos que ya conocía. Dennett sostiene que la verdad de la premisa primera es incompatible con la de la segunda. Si Mary conoce toda la información física sobre la visión de los colores, será capaz de anticipar qué efectos tendrá sobre su sistema nervioso cada color particular. La única tarea que le queda es figurarse un modo de identificar esos efectos desde el interior. Parece, no obstante, que la posición de Dennett contiene un *non sequitur*. Mary podría ser capaz de anticipar e identificar los efectos de los colores sobre su sistema nervioso sin ser capaz de anticipar las propiedades fenomenológicas de su experiencia de los colores.

b) La réplica en términos de capacidades

Lawrence Nemirow (1980, 1990) y David Lewis (1983, 1988) sostienen que Mary sólo gana saber-cómo, no saber-qué. Lo que Mary adquiere cuando experimenta por vez primera una sensación de rojo no es conocimiento proposicional, sino meramente una capacidad recognitiva. Así el argumento del conocimiento comete la *falacia de equivocidad*. Cuan-

do en la primera premisa se afirma que Mary conoce todo lo físico que hay que conocer sobre la visión de los colores, «conocer» se usa en el sentido de conocimiento proposicional. El conocimiento es en este caso un asunto de haber dominado un conjunto de proposiciones o de información. Pero cuando se dice en la segunda premisa que Mary conoce algo nuevo sobre el color rojo, «conocer» hace referencia al conocimiento práctico. El conocimiento es ahora un asunto de tener una representación mental de la rojez o de ser capaz de hacer ciertas discriminaciones sensoriales, de ser capaz de recordar e imaginar, de poder reconocer la misma experiencia si se presenta de nuevo.

¿Pero es plausible esta réplica? Ciertamente, parte de lo que Mary gana es saber-cómo, pero no es eso todo lo que gana. Nos sentimos inclinados a decir que hay un hecho que ella aprehende sólo después de su liberación, un hecho sobre cómo aparece el rojo fenoménico. Después de todo, saber *cómo* se siente algo es saber *que* se siente de un cierto modo. Así el antifisicista parece tener razón cuando afirma que Mary entra en posesión de cierta información distintiva. Pero, aun admitiendo esto, cabe preguntarse si Mary tan sólo llega a conocer *de un modo nuevo* hechos o proposiciones que ya conocía.

c) ¿Viejos conocimientos en odres nuevos?

Algunos filósofos, como Paul Churchland (1985, 1990) y M. Tye (1986), han mantenido que Mary gana nuevo conocimiento sólo en el sentido de que adquiere un nuevo modo de acceso a proposiciones y hechos que ya conocía. Mary conoce ahora directamente por introspección lo que antes conocía sólo indirectamente por inferencia. Ahora es capaz de representar los viejos hechos que ya conocía usando un sistema de representación biológico y probablemente prelingüístico distinto de las capacidades lingüísticas que tuvo que usar para representar esos hechos en el pasado. Así podemos decir que Mary está en un nuevo estado epistémico debido a esas diferencias en el *modo de acceso* o *sistema de representación*.

Lewis (1988) ha criticado esta propuesta considerándola inadecuada para dar cuenta del argumento. En su opinión, sus propugnadores hacen que el sentido en el que Mary obtiene nuevo conocimiento al experimentar la sensación de rojo no sea diferente de aquél en el que obtiene nuevo conocimiento sobre su cerebro cuando aprende ruso o urdu y así adquiere un nuevo sistema de representación.

d) Proposiciones de grano fino

Si rechazamos tanto la réplica en términos de capacidades como la réplica en términos de un nuevo sistema de representación, aún es posible intentar bloquear el argumento del conocimiento cuestionando lo que ha

de contar como una proposición. Las proposiciones, individuadas como entidades no estructuradas, pueden entenderse como funciones de mundos posibles a valores de verdad (o, equivalentemente, como conjuntos de mundos posibles). Bajo tal modo de individuación, la proposición de que el agua hierve a 100ºC es la misma que la proposición de que el compuesto H_2O hierve a 100ºC, puesto que ambas son verdaderas en los mismos mundos posibles. Bajo un modo de individuación más fino, las proposiciones pueden ser consideradas entidades *estructuradas* compuestas de *conceptos* que deben a su vez encajar para que dos proposiciones sean idénticas. ¿Bajo qué modo de individuación aprende Mary una nueva proposición? ¿Bajo el modo de individuación de grano grueso o bajo el modo de individuación de grano fino? La segunda opción proporciona otra vía de escape a la conclusión del argumento del conocimiento. Es la vía favorecida por Brian Loar (1990), y Robert van Gulick (1993). Loar arguye que Mary adquiere un nuevo concepto que entra en su repertorio cognitivo sobre la base de sus capacidades discriminativas recientemente adquiridas. Y, al usar ese nuevo concepto, es capaz de captar la verdad de nuevas proposiciones. Pero la adición de esas nuevas proposiciones a su repertorio no representa un problema para el fisicista, ya que la *propiedad* a la que se refiere su nuevo *concepto* puede ser sólo una propiedad a la que ya se refería en el pasado mediante el uso de un concepto puramente físico.

Es cierto que cuando se trata de un género natural el concepto fenoménico y el concepto físico-funcional pueden divergir en rol cognitivo aunque converjan en la propiedad que denotan. Aunque el concepto de un líquido incoloro, inodoro, insípido, etc., puede desempeñar un papel cognitivo distinto que el del concepto de H_2O, ambos conceptos introducen la misma propiedad, para utilizar la terminología de Loar. Pero el argumento no es trasladable al caso de un color. El concepto de rojo no puede identificarse con ninguna propiedad física subyacente —como una cierta longitud de onda—. Un argumento de «realización variable» pone de manifiesto la imposibilidad de tal identificación. Supongamos que descubriéramos que las propiedades físicas de los objetos variasen realmente pero que la variación fuese compensada en nuestros órganos visuales. No diríamos entonces que los objetos han variado de color. Diríamos que los objetos parecen de tal y cual color y por ello son de tal y cual color. Dicho de otro modo, las cualidades secundarias son independientes de su fundamento físico. Son definidas disposicionalmente, como hemos visto más arriba. Los géneros naturales, en cambio, son definibles en términos de «esencias reales» lockeanas. De ahí que la réplica de Loar funcione sólo para géneros naturales y no para *qualia*.

3. *La elusividad de la conciencia*

a) El agnosticismo de Thomas Nagel

En su famoso ensayo «What Is It Like to Be a Bat?», Thomas Nagel (1974) se ha mostrado escéptico acerca de la posibilidad de alcanzar el nivel subjetivo de la fenomenología a partir del nivel objetivo de la fisiología. Nagel comienza afirmando que ningún análisis reduccionista de los fenómenos mentales puede dar cuenta de la conciencia. «Sin la conciencia», escribe, «el problema mente-cuerpo sería mucho menos interesante. Con la conciencia parece desesperado» (*o. c.*, 159). El hecho de que un organismo tiene estados mentales conscientes significa que «hay algo que es cómo *ser* ese organismo» (*o. c.*, 160). Para que el fisicismo tenga éxito, ese carácter subjetivo de la experiencia, ese conjunto de rasgos fenomenológicos que determinan cómo es ser ese organismo, debe recibir una explicación física. Pero el carácter subjetivo de la experiencia se escapa a cualquier explicación física porque «está esencialmente conectado con un único punto de vista y parece inevitable que una teoría física, objetiva, abandone ese punto de vista» (*o. c.*, 160).

Para ilustrar este abismo entre los tipos de comprensión subjetiva y objetiva, Nagel se pregunta si podemos imaginar cómo es ser un murciélago que se traslada por ecolocalización. Su respuesta es que ninguna cantidad de información física nos permite captar cómo es ser tal creatura. Nuestra propia experiencia proporciona el material básico para nuestra imaginación y así su ámbito es restringido. Puedo imaginarme lo que sería para *mí* comportarme como un murciélago, pero no puedo imaginar lo que es para un *murciélago* ser un murciélago. Hay, pues, hechos que escapan a la comprensión que nos permiten nuestros limitados esquemas conceptuales humanos.

Cuando Nagel afirma que los hechos fenomenológicos son accesibles sólo desde un punto de vista, no entiende un punto de vista como algo privado a un individuo. No se trata de una reedición del escepticismo sobre otras mentes. A este respecto afirma:

> No estoy refiriéndome aquí a la pretendida privacidad de la experiencia para su poseedor. El punto de vista en cuestión no es accesible sólo a un único individuo. Es más bien un *tipo*... una persona puede conocer o decir de otra cuál es la cualidad de la experiencia de la otra. Son subjetivos, sin embargo, en el sentido de que incluso esta adscripción objetiva de experiencia sólo es posible para alguien suficientemente similar al objeto de adscripción para poder adoptar su punto de vista (*o. c.*, 163).

Así un punto de vista es algo compartible por muchos individuos en virtud de tener sistemas perceptivos semejantes. Pero un punto de vista constituye una limitación a lo que es concebible por un individuo. Ahora

bien, si la experiencia no tiene una naturaleza objetiva, si los hechos de experiencia son sólo accesibles desde un punto de vista, su carácter no puede revelarse en el dominio objetivo, accesible desde muchos puntos de vista, de las operaciones físicas de un organismo. Cuanto más nos despeguemos de un punto de vista específico en la dirección de una mayor objetividad, más nos alejaremos de la naturaleza real de los fenómenos experienciales.

Nagel cree que esto incide directamente en el problema mente-cuerpo: «si los hechos de la experiencia... son accesibles sólo desde un solo punto de vista, entonces es un misterio cómo podría revelarse el verdadero carácter de las experiencias en la operación física» (*o. c.*, 163) de un organismo. Nagel no concluye que el fisicismo debe ser falso. La conclusión apropiada es en su opinión que el fisicismo es una posición que no podemos entender. En (1986, 47) expresa su agnosticismo en estos términos: «No tenemos en la actualidad concepción ninguna de cómo un solo evento o cosa podría tener a la vez aspectos físicos y fenomenológicos, ni de cómo podrían estar relacionados si los tuviera». Nagel se inclina por una *teoría de aspecto dual* en la línea de Spinoza, según la cual los fenómenos mentales son los aspectos subjetivos de estados que admiten también descripción física.

b) El naturalismo no-constructivo de Colin McGinn

Una posición hasta cierto punto similar ha sido defendida recientemente por McGinn (1989, 1991). McGinn busca una tercera vía entre la Escila de las explicaciones no-naturalistas y el Caribdis del naturalismo constructivo. Para el no-naturalista el asiento de la conciencia está en alguna substancia ultraterrena. El dualismo cartesiano es el paradigma de esta concepción. McGinn la rechaza sobre la base de la consideración naturalista de que «la conciencia... debe ser un fenómeno natural que surge naturalmente de ciertas organizaciones de la materia» (Mc Ginn, 1989, 353). Para el naturalista constructivo, la explicación de cómo es que el cerebro material es el asiento de la conciencia fenoménica se encuentra en alguna propiedad natural que ejemplifican los cerebros. El funcionalismo y el materialismo del estado central son dos posiciones paradigmáticas al respecto. Pero el argumento de los qualia ausentes muestra que el mero hecho de que el cerebro ejemplifique propiedades funcionales no nos aclara por qué tenemos conciencia. Y los argumentos de Nagel y Jackson muestran que el materialismo del estado central fracasa.

La vía media de McGinn es el *naturalismo no-constructivo.* El cerebro es el asiento de la conciencia en virtud de ciertas propiedades suyas; pero qué sean esas propiedades y cómo dan lugar a la conciencia fenoménica está más allá de nuestra captación cognitiva. Estamos en una situación de *clausura cognitiva* con respecto a tales asuntos.

Supongamos que tuviéramos una teoría psicofísica T explicativa del nexo causal existente entre la estructura fisiológica del cerebro de los murciélagos y el tipo de experiencias (M-experiencias, para abreviar) de esas creaturas. Llamemos P a la propiedad explicativa que liga M-experiencias con el cerebro del murciélago. Entonces captar T habría de conferirnos una captación de la naturaleza de esas experiencias. Esto es así porque captar la naturaleza de las M-experiencias nos llevaría a captar el *carácter*, la forma subjetiva, de las M-experiencias, pues estaríamos en posesión del mismo tipo de comprensión que tendríamos de nuestras propias experiencias si tuviésemos la teoría psicofísica correcta de ellas. Pero ahora nos encontramos con un dilema. O bien podemos captar T, en cuyo caso la propiedad M se nos vuelve abierta. O no podemos captar T, simplemente porque M no nos está abierta. Pero si T no logra conferir una captación de la naturaleza y el carácter de esas experiencias, estamos en la absurda posición de entender una teoría que hace perfectamente inteligible la relación entre un conjunto de fenómenos que no entendemos; esto es, estamos en la absurda situación de entender T sin entender el concepto M que aparece en ella.

La conclusión de McGinn es escéptica: «Existe alguna propiedad del cerebro que da cuenta naturalísticamente de la conciencia», pero «estamos cognitivamente cerrados a esa propiedad» (Mc Ginn, 1989, 352) en el sentido de que los procedimientos de formación de conceptos de los que disponemos no pueden extenderse a su captación, del mismo modo que un niño de cinco años es constitucionalmente incapaz de entender la Teoría de la Relatividad. Este escepticismo de McGinn contrasta con el agnosticismo de Nagel, quien, comentando la tesis de la *clausura cognitiva* de aquél escribe: «Aunque pudiera estar en lo cierto, creo que su pesimismo es prematuro» (Nagel, 1993, 40).

En *Consciousnes Reconsidered*, Owen Flanagan (1992) se embarca en una sistemática defensa del naturalismo constructivo. Contra el anticonstructivismo de McGinn y el agnosticismo de Nagel, Flanagan se muestra optimista acerca de nuestra capacidad de entender la relación entre conciencia y cerebro. Opina que la existencia de la conciencia en el mundo natural puede hacerse inteligible mediante los resultados combinados de la neurociencia, la psicología cognitiva y la filosofía de la mente. Contra el naturalismo eliminativo de Paul y Patricia Churchland, mantiene que el concepto de conciencia cualitativa es necesario, al menos al comienzo de la investigación, para delimitar qué es lo que es necesario explicar. Tenemos que usar, aunque sean revisables, nuestros modos ordinarios de taxonomizar la subjetividad a fin de atacar el problema de la conciencia.

c) Searle y la subjetividad ontológica de la conciencia

John R. Searle se alinea con las posiciones de Nagel y McGinn en su último libro *The Rediscovery of Mind* (1992). En su opinión las variedades del materialismo que hay en el mercado filosófico son erróneas en un punto crucial: todas ellas dejan fuera la conciencia. Hay, según Searle, una diferencia fundamental entre los fenómenos conscientes y los fenómenos comportamentales o fisiológicos que hace imposible la reducción de los primeros a los segundos: la conciencia es *ontológicamente subjetiva*. Sus rasgos esenciales no pueden descubrirse enteramente desde un punto de vista externo, en tercera persona. Es indispensable el punto de vista de la primera persona a fin de captar cómo es un estado mental concreto para un sujeto.

Esto no quiere decir que Searle abrace el dualismo. Enfáticamente rechaza tanto el dualismo de substancias cartesiano como el dualismo de propiedades —la concepción según la cual, aunque no hay un alma distinta del cuerpo, los fenómenos mentales involucran propiedades del sujeto que no son físicas—. Para Searle, la conciencia es una propiedad física del cerebro, a pesar de su subjetividad, aunque es irreducible a cualesquiera otras propiedades físicas. En su opinión, los materialistas reduccionistas asumen injustificadamente que, si los fenómenos mentales no pueden explicarse en términos psicofísicos, la única alternativa restante es el dualismo. Pero debemos rechazar el supuesto de que debe haber uno o dos géneros últimos de cosas o propiedades. Las teorías materialistas confunden además dos sentidos de la distinción subjetivo/objetivo: el sentido epistemológico y el sentido ontológico. Epistémicamente, la objetividad es un *desideratum* metodológico de la ciencia que nos impulsa a buscar la máxima independencia de valores, prejuicios, puntos de vista y emociones. Ontológicamente, la distinción señala diferentes categorías de la realidad empírica. La subjetividad ontológica de la conciencia debe ser reconocida como tal y descrita en términos epistémicamente objetivos. ¿Cómo alcanzar una concepción científica epistémicamente objetiva de un mundo que contiene los hechos ontológicamente subjetivos de la conciencia?

Searle propugna la adopción de un *naturalismo biológico* que combina la irreducible subjetividad de la conciencia con el rechazo de la dicotomía entre lo mental y lo físico. Para él, la conciencia es un rasgo biológico de los seres humanos y de ciertos animales que es causado por procesos neurobiológicos. Es una propiedad de segundo orden o emergente del cerebro. Es una propiedad mental, *y por tanto física*, del cerebro en el sentido en el que la liquidez es una propiedad emergente de sistemas de moléculas.

En su reseña del libro, Tom Nagel (1993) observa que si aceptamos que la objetividad ontológica es uno de los rasgos definitorios de los fe-

nómenos físicos y si aceptamos que la conciencia es ontológicamente subjetiva, entonces Searle carece de justificación para llamar «físicos» a los rasgos irreducíblemente subjetivos del cerebro. De hecho, piensa Nagel, Searle no ha logrado diseñar una postura que se diferencie genuinamente del dualismo de propiedades. Si la distinción ontológica entre lo subjetivo y lo objetivo delimita diferentes categorías de la realidad empírica, decir que el universo contiene un componente físico irreduciblemente subjetivo es formular «una tesis esencialmente dualista en un lenguaje que expresa una fuerte aversión al dualismo» (*o. c.*, 40). Por otro lado, Nagel piensa que la comparación de la relación entre la conciencia y la conducta de las neuronas con la relación entre la liquidez y la conducta de las moléculas de H_2O es desorientadora. Searle reconoce que podemos representar figurativamente desde fuera la relación necesaria entre los niveles macroscópicos y microscópicos del agua, pero no podemos hacer esto con la subjetividad, que tenemos que imaginar siempre desde el interior. Nagel está de acuerdo, pero vuelve a insistir en su conclusión agnóstica: «Creo que esto significa que no entendemos realmente la afirmación de que los estados mentales son estados del cerebro. Somos aún incapaces de formar una concepción de *cómo* surge la conciencia de la materia, aunque estemos seguros de que lo hace» (*o. c.*, 40).

BIBLIOGRAFÍA

Block, N. (1978), «Troubles with Functionalism», en W. Savage (ed.), *Perception and Cognition: Minnesota Studies in the Philosophy of Science*, vol. IX, University of Minnesota Press, Minneapolis; reimp. en Block, 1980.

Block, N. (ed.) (1980), *Readings in the Philosophy of Psychology*, 2 vols., Harvard University Press, Cambridge, Mass.

Block, N. (1990), «Inverted Earth», en Tomberlin, 1990.

Block, N. y Fodor, J. (1974), «What Psychological States Are Not»: *Philosophical Review*, *81*, 159-81; reimp. en Block, 1980.

Churchland, P. M. (1984), *Matter and Consciousness*, MIT Press, Cambridge, Mass. V.e.: M. N. Mizraji, *Materia y conciencia*, Gedisa, Barcelona, 1992.

Churchland, P. M. (1985), «Reduction, Qualia, and the Direct Introspection of Brain States»: *Journal of Philosophy*, *82*, 8-28.

Churchland, P. M. (1990), «Knowing Qualia: A reply to Jackson», en Id., *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*, MIT Press, Cambridge, Mass.

Davies, M. y Humphreys, G. W. (eds.) (1993), *Consciousness: Psychological and Philosophical Essays*, Blackwell, Oxford.

Dennett, D. (1988), «Quining Qualia», en A. Marcel y E. Bisiach (eds.), *Consciousness in Contemporary Science*, Oxford University Press, Oxford; reimp. en Lycan, 1990.

Dennett, D. (1991), *Consciousness Explained*, Allen Lane-The Penguin Press, London.

Flanagan, O. (1992), *Consciousness Reconsidered*, MIT Press, Cambridge, Mass.
García Suárez, A. (1976), *La lógica de la experiencia de Wittgenstein y el problema del lenguaje privado*, Tecnos, Madrid.
Gregory, R. L. (1974), *Concepts and Mechanisms of Perception*, Duckworth, London.
Harman, G. (1989), «Some Philosophical Issues in Cognitive Science: Qualia, Intentionality, and the Mind-Body Problem», en M. I. Posner (ed.), *Foundations of Cognitive Science*, MIT Press, Cambridge, Mass.
Harman, G. (1990), «The Intrinsic Quality of Experience», en Tomberlin, 1990.
Jackson, F. (1982), «Epiphenomenal Qualia»: *Philosophical Quarterly, 32*, 127-36; reimp. en Lycan, 1990.
Jackson, F. (1986), «What Mary Didn't Know»: *Journal of Philosophy*, *83*, 291-5; reimp. en D. Rosenthal (ed.), *The Nature of Mind*, Oxford University Press, Oxford, 1991.
Kripke, S. (1971), «Identity and Necessity», en M. K. Munitz (ed.), *Identity and Individuation*, New York University Press. V.e.: M. Valdés, *Identidad y necesidad*, Instituto de Investigaciones Filosóficas-UNAM, México, 1978.
Kripke, S. (1972), «Meaning and Necessity», en D. Davidson y G. Harman (eds.), *Semantics of Natural Language*, D. Reidel, Dordrecht; reimp. en forma de libro con el mismo título y un apéndice por Harvard University Press, Cambridge, Mass., 1980. V.e.: M. Valdés, *El nombrar y la necesidad*, Instituto de Investigaciones Filosóficas-UNAM, México, 1985.
Levin, M. (1975), «Kripke's Argument against the Identity Theory»: *Journal of Philosophy*, 72, 149-67.
Levine, J. (1993), «On Leaving Out What It's Like», en Davies y Humphreys, 1993.
Lewis, D. (1980), «Mad Pain and Martian Pain», en Block, 1980.
Lewis, D. (1983), «Postscript to "Mad Pain and Martian Pain"», en Id., *Philosophical Papers*, vol. I, Oxford University Press, Oxford.
Lewis, D. (1988), «What Experience Teaches», *Actas de la Russellian Society de la Universidad de Sidney*; reimp. en Lycan, 1990.
Loar, B. (1990), «Phenomenal Properties», en Tomberlin, 1990.
Locke, J. (1690), *An Essay concerning Human Understanding*, en P. H. Nidditch, (ed.), Clarendon, Oxford, 1975. V.e.: O'Gorman, *Ensayo sobre el entendimiento humano*, FCE, México, 1956.
Lycan, W. G. (1974), «Kripke and the Materialists»: *Journal of Philosophy*, *71*, 677-89.
Lycan, W. G. (1987), *Consciousness*, MIT Press, Cambridge, Mass.
Lycan, W. G. (ed.) (1990), *Mind and Cognition: A Reader*, Blackwell, Oxford.
McGinn, C. (1983), *The Subjective View: Secondary Qualities and Indexical Thoughts*, Clarendon Press, Oxford.
McGinn, C. (1989), «Can We Solve the Mind-Body Problem?»: *Mind*, 93, 349-64; reimpr. en Mc Ginn, 1991.
McGinn, C. (1991), *The Problem of Consciousness*, Blackwell, Oxford.
Nagel, T. (1974), «What Is It Like to Be a Bat?»: *Philosophical Review*, 83, 435-50, reimp. en Block, 1980.
Nagel, T. (1986), *The View from Nowhere*, Oxford University Press, Oxford.

Nagel, T. (1993), «Recensión de J. Searle», *The Rediscovery of Mind: The New York Review of Books*, XL, 5, 4 de marzo, 37-41.
Nemirow, L. (1980), «Recensión de T. Nagel, *Mortal Questions*»: *Philosophical Review*, *89*, 475-6.
Nemirow, L. (1990), «Physicalism and the Cognitive Role of Acquaintance», en Lycan, 1990.
Sanfélix Vidarte, V. (1991), «Panorama actual de la filosofía analítica de la mente: funcionalismo y experiencia», en M. Torrevejano (ed.), *Filosofía analítica hoy*, Universidad de Santiago de Compostela, Santiago de Compostela.
Searle, J. R. (1980), «Minds, Brains, and Programs»: *Behavioral and Brain Sciences*, *3*, 417-24.
Searle, J. R. (1983), Intentionality: *An Essay in the Philosophy of Mind*, Cambridge University Press, Cambridge.
Searle, J. R. (1984), *Minds, Brains, and Science: The 1984 Reith Lectures*, Harvard University Press, Cambridge, Mass. V.e.: L. M. Valdés Villanueva, *Mentes, cerebros y ciencia*, Cátedra, Madrid, 1985.
Searle, J. R. (1992), *The Rediscovery of Mind*, MIT Press, Cambridge, Mass.
Shoemaker, S. (1975a), «Phenomenal Similarity»: *Crítica*, 7, 3-34; reimp. en Shoemaker, 1984.
Shoemaker, S. (1975b), «Functionalism and Qualia»: *Philosophical Studies*, *27*, 291-315; reimp. en Block, 1980 y en Shoemaker, 1984.
Shoemaker, S. (1980), «Absent Qualia Are Impossible: —a Reply to Block»: *Philosophical Review*, 90, 581-99; reimp. en Shoemaker, 1984.
Shoemaker, S. (1981), «The Inverted Spectrum»: *Journal of Philosophy*, *74*, 357-81, reimp. en Shoemaker, 1984.
Shoemaker, S. (1984), *Identity, Cause, and Mind: Philosophical Essays*, Cambridge University Press, Cambridge.
Shoemaker, S. (1991), «Qualia and Consciousness»: *Mind*, *100*, 507-24.
Tomberlin, J. (ed.) (1990), *Philosophical Perspectives*, 4: *Action Theory and Philosophy of Mind*, Ridgview Publishing, Atascadero, CA.
Tye, M. (1986), «The Subjective Qualities of Experience»: *Mind*, *95*, 1-17.
Gulick, R. (1993), «Understanding the Phenomenal Mind: Are We All Just Armadillos?», en Davies y Humphreys.
Wittgenstein, L. (1953), *Philosophical Investigations*, Blackwell, Oxford. V.e.: A. G. Suárez y U. Moulines, *Investigaciones filosóficas*, Crítica/UNAM, Barcelona y México, 1988.

CONCIENCIA

Enrique Villanueva

I. LA REVOLUCIÓN CARTESIANA

Descartes llevó a cabo una revolución en el pensamiento metafísico occidental cuando movió la propiedad de la conciencia ubicándola como la propiedad central esencial de la mente, una propiedad que toda otra propiedad debe tener para poder ser calificada como una propiedad mental[1]. El pensamiento, por ejemplo, no puede calificar como una propiedad mental a menos que sea pensamiento consciente. Tener mente es tener conciencia y viceversa. La esencia del pensamiento es la conciencia. La conciencia resulta, de esta manera, algo esencial o intrínseco a toda propiedad mental.

Descartes, como muchos otros pensadores, nunca ofrece un análisis de la conciencia, pues le parece que es una propiedad cuya naturaleza se hace manifiesta en la experiencia. Basta experimentar un pensamiento, percepción, recuerdo, deseo, etc., para que nos demos cuenta del mismo, para que nos percatemos de su contenido pensado, su contenido sentiente, etc. La conciencia toda se manifiesta en la experiencia y la experiencia no revela ninguna estructura o composición en la propiedad de la conciencia.

Otra nota esencial que Descartes confiere a la conciencia es la de ser

1. Descartes (1897). Consúltense especialmente sus *Meditaciones Metafísicas* y las *Objeciones y respuestas*. En las *Cuartas respuestas* dice: «Ningún pensamiento puede existir en nosotros del que no seamos conscientes en el mismo momento en que existe en nosotros». Igualmente dice: «La palabra "pensamiento" se aplica a todo lo que existe en nosotros de tal manera que somos inmediatamente conscientes de ello».

Descartes se aparta de la concepción clásica de Platón y Aristóteles introduciendo la perspectiva de la subjetividad o del mentalismo.

un factor unificador que reúne los estados mentales en un único sitio mental que denomina ego. Este ego lo caracteriza adicionalmente como un individuo simple cuya esencia es el pensamiento consciente. La personeidad toda consiste en este ego unificador. Pero, además, ser persona de esa manera es una cuestión de todo o nada para Descartes [2].

La conciencia es la base de la certeza y la certeza es la base de la racionalidad, pues solamente la primera puede enfrentar exitosamente el desafío que le hace el escéptico metafísico radical en el sentido de que podemos estar absolutamente engañados respecto del mundo, de nosotros mismos y de nuestros propios pensamientos. Descartes cree que encuentra un principio epistemológico fundamental [3] que le permite enfrentar a ese escéptico radical y asegurar la realidad del mundo, la de su identidad mental que denomina ego, y la de cada propiedad mental particular. Ese principio epistemológico fundamental descansa en la transparencia de la conciencia tal como la concibe Descartes. Toda la racionalidad de que son capaces las personas descansa en esa verificación constante que lleva a cabo la actividad ininterrumpida de la conciencia. Descartes compara la conciencia con una luz que ilumina a los objetos que caen bajo su haz y solamente es racional creer en esos objetos en los términos que la propia conciencia establece.

La propiedad de la conciencia es la propiedad esencial de la mente y la mente es radicalmente distinta del cuerpo; en consecuencia, la conciencia no es una propiedad natural del tipo de las propiedades que explica la ciencia física, por ejemplo. Para Descartes la conciencia es una propiedad irreductiblemente no-natural, no-física.

Debido a que la conciencia sólo se da en la experiencia y no hay análisis de la misma, todo lo que sabemos de ella está dado por esta metafísica a priori y nada más podemos saber acerca de la misma excepto lo que nos da este contacto intuitivo en la experiencia [4]. Este es el verificacionismo mentalista que introduce Descartes.

En suma, para Descartes la conciencia es una propiedad transparente, inmediata, con privilegio epistemológico, simple, inanalizable y no-natural.

2. Esta tesis del *ego* simple, unificador, se ve fuertemente controvertida por los estudiosos de la ciencia cognitiva para quienes la mente humana consiste en una serie de módulos que operan en coordinación. La concepción modular de la mente está expuesta en Fodor (1983).

3. Sobre esta interpretación consúltense mis trabajos sobre Descartes incluídos en *Ensayos de Historia Filosófica*, UNAM, México, 1988 y mi libro *Las Personas*, en prensa. En su tesis de la certeza Descartes confunde la conciencia fenomenal con la conciencia representacional; así, por ejemplo, dice que el contenido fenomenal es verdadero.

4. Descartes inaugura esta manera de pensar, pero su interaccionismo causal no implica esta tesis.

II. CONTRA CARTESIANOS: LEIBNIZ Y KANT

Esta concepción verificacionista de la conciencia la rechazan ambos, Leibniz y Kant. En varios lugares Leibniz[5] habla de las percepciones y los pensamientos inconscientes abriendo con ello la puerta para una investigación del pensamiento y otras propiedades mentales que no dependa de la experiencia humana consciente al tiempo que permite que dicha investigación tenga un carácter empírico similar a la de la física. Leibniz rompe la igualdad (mente = conciencia) y con ello abre la posibilidad de estudiar empíricamente los procesos mentales sin recurrir, por ejemplo, a la conciencia o a la introspección. Su teoría paralelista de la relación entre la mente y el cuerpo se traduce en la tesis empírica de que nada hay en la mente que no se encuentre en el cuerpo, de manera que toda la investigacion de los estados y procesos mentales se puede llevar a cabo en el cuerpo.

Kant sigue a Leibniz en esta posición anti-cartesiana y al llevar a cabo su investigación sobre las funciones trascendentales del pensamiento teórico y práctico elimina el requisito de la verificación consciente[6]. Todo el complejo mecanismo del pensamiento corre sin que la conciencia tenga que supervisarlo. Kant lleva a cabo una investigación que constituye una teoría compleja del pensamiento que puede reconstruirse desde el punto de vista de la teoría funcionalista actual[7].

III. POSTCARTESIANOS

En la segunda mitad del siglo pasado varios autores de origen cartesiano ofrecen teorías de la conciencia, entre los más importantes se encuentran Franz Brentano y William James. El primero sostiene[8] que la conciencia es una propiedad que tiene como condición necesaria la intencionalidad. La intencionalidad consiste en una especie de transitividad, de estar dirigido hacia algo y a ese algo se le denomina el objeto intencional. Para Brentano, toda conciencia es intencional y toda intencionalidad es consciente o transparente. Un estado es consciente, en parte, porque es acerca de él mismo, y puesto que todo estado mental es consciente (es acerca de él mismo), se sigue que la propiedad de la conciencia es intrínseca a todo estado mental. Toda conciencia tiene un objeto intencional el cual

5. Consúltense especialmente los *Nouveaux Essais Sur Le Entendement Humain*.

6. Consúltese la *Crítica de la Razón Pura*. Por razones independientes, Kant se expresa en formas que parecen preservar el carácter epistemológico en su investigación del pensamiento y en esa misma medida parece quedar cautivo del error cartesiano. Sin embargo, véase la lectura que hace (1990).

7. Kitcher (1990) ha desarrollado esa interpretación en parte.

8. Brentano (1882). Su influencia en el pensamiento alemán y francés es determinante, lo cual explica la infecundidad de todo ese pensamiento cartesiano; recuérdese el caso notorio de Husserl. En nuestros días Armstrong (1968) y Rosenthal (1986) rescatan de una manera diferente —según se apreciará más adelante— la tesis de Brentano.

determina el contenido de esa conciencia. En consecuencia, no puede haber una investigación de la conciencia que no tome en cuenta esta cautividad por la intencionalidad. Paralelamente, toda investigación de la intencionalidad exige una investigación de la conciencia, pues toda propiedad intencional incluye la propiedad de la conciencia. Brentano anticipa parcialmente el análisis contemporáneo según el cual la conciencia resulta de la incidencia de un pensamiento en otro pensamiento. Como se puede inferir de la distinción que introduciremos más adelante, Brentano piensa en términos de conciencia proposicional y no en términos de conciencia fenomenal.

Esta tesis de la conciencia de Brentano como algo intrínseco a todo estado mental y como inseparable de la intencionalidad reinstala con toda su fuerza la tesis mentalista cartesiana y resulta altamente controvertida en nuestros días. Muchos la rechazan con una gran variedad de argumentos a la vez empíricos y *a priori* en la investigación empírica acerca de la conciencia. Muchos otros la defienden; así por ejemplo, Searle[9], quien la defiende parcialmente para criticar principalmente la investigación cognitiva actual como una investigación que deja de lado la propiedad misma de la conciencia.

William James[10] reconoce la necesidad de una investigación empírica de la conciencia, pero siempre dentro del presupuesto cartesiano de que conciencia es experiencia, de que en el caso de la conciencia *esse est sentire*. James afirma la conexión intrínseca entre la conciencia y el carácter sensorial. Las varias tesis de la conciencia que defiende James se resienten del conflicto entre el carácter empírico y el presupuesto cartesiano que nunca abandona por completo. Este presupuesto limita severamente muchas de sus importantes aportaciones, entre las cuales se encuentran el carácter activo de la conciencia, el rechazo de la introspección y su diferencia con el contenido de los estados mentales.

IV. LA INVESTIGACIÓN DE LA CONCIENCIA EN NUESTROS DÍAS

La reciente investigación de la conciencia establece su carácter empírico, no-experiencial y natural. Toda determinación metafísica de la conciencia debe partir de conocimientos y todo conocimiento de la conciencia proviene de la experimentación empírica. Sin embargo, no es necesario tener experiencia de las propiedades de la conciencia. Finalmente, la naturaleza de la conciencia no difiere ontológicamente de la naturaleza de los objetos que investiga la demás ciencia empírica; la propiedad de la conciencia no constituye ningún reino ontológico peculiar.

9. Searle (1991).
10. James (1890).

V. CIENCIA DE LA CONCIENCIA Y FILOSOFÍA DE LA CONCIENCIA

Una cuestión central debatida hasta nuestros días inquiere si la propiedad de la conciencia es apropiada para un tratamiento empírico, científico, del tipo de la física o la neurociencia. Algunos piensan que hay por lo menos algún tipo de conciencia que resulta reticente al tratamiento científico. A este tipo de conciencia se le ha llamado la conciencia fenomenal y responde a lo que se da en cada experiencia y de la cual cada persona puede ser testigo. En nuestros días se habla del carácter subjetivo, experiencial, perspectivo, «el como qué es...» o del tipo de conciencia que es la conciencia fenomenal[11]. Han sido algunos filósofos los que han insistido en una elucidación de la conciencia y en el fracaso de las teorías y modelos cognitivos de poder ofrecer un tratamiento adecuado de esta propiedad. Podemos aislar una cuestión filosófica de la conciencia, a saber, qué es este aparecer, experimentar que resulta tan manifiestamente diferente de las cosas físicas; ¿es posible proveer una elucidación de este aparecer que se condiga con la naturaleza de las propiedades que apreciamos en las cosas físicas?

En el campo científico la neurociencia ha podido localizar las funciones conscientes con gran precisión y los científicos cognitivos han elaborado modelos que establecen hipótesis acerca de la conciencia y se confirman mediante experimentos, según podremos ilustrar más adelante. Podemos poner la cuestión científica como una de explicar la propiedad de la conciencia, de controlar su ocurrencia, de conocer su función y, en consecuencia, de poder duplicarla en mecanismos diferentes al cuerpo humano.

En nuestros días la disputa científica se establece entre dos posiciones, a saber, los que mantienen un escepticismo de poder elaborar una teoría de la naturaleza de la conciencia que resulte explicativa de otras propiedades mentales así como de hechos del mundo y los que mantienen un optimismo de que es posible alcanzar tal conocimiento en forma gradual, comenzando por tipos específicos de conciencia hasta alcanzar otros más complejos y finalmente la propiedad toda de la conciencia.

Otra manera de expresar esta oposición consiste en afirmar que es posible tener una teoría del pensamiento o contenido de las propiedades mentales pero no es posible tener una teoría de la conciencia, pues esta última implica tener una teoría de la mente toda, mientras que la primera solamente trata de una parte de la mente.

En toda esta disputa resulta crucial determinar si la propiedad de la intencionalidad implica la propiedad de la conciencia o si es la propiedad

11. T. Nagel (1979 y 1986) introdujo el problema. N. Block (1993) elabora la distinción entre este tipo de conciencia fenomenal y otro tipo al que llama conciencia-acceso. Esta distinción aparece más adelante.

de la conciencia la que implica la propiedad de la intencionalidad o si finalmente se trata de propiedades independientes [12]. La intencionalidad tiene dos notas independientes, a saber, la transitividad u objectualidad, el carácter «acerca de» y la perspectividad o aspectualidad, el hecho de que se dé en aspectos o escorzos. Esta segunda nota resulta la más recalcitrante, mientras que la primera es controvertida (especialmente en el caso de la conciencia fenomenal que precisamente no consiste en ser acerca de nada, sino en su propio aparecer o manifestarse).

Otra dificultad adicional concierne si el carácter sensorial le es esencial o intrínseco a la propiedad de la conciencia.

Hay que observar que la disputa filosófica de la conciencia no encuentra solución [13] y esto se debe en gran parte a la falta de conocimiento acerca de la propiedad de la conciencia. En lugar de conocimientos hay argumentos filosóficos *a priori* que se mueven en un nivel en el que presuponen que la propiedad de la conciencia tiene las notas que el lenguaje ordinario le atribuye y es esta presuposición la que exhibe una falta de fundamento y tal vez de coherencia [14]. Es necesario determinar cuáles de las notas que expresa el lenguaje ordinario pertenecen a la conciencia y cuáles no le pertenecen.

VI. METAFÍSICA Y EPISTEMOLOGÍA DE LA CONCIENCIA

Un grave error de la investigación sobre la naturaleza de la conciencia consiste en introducir cuestiones epistemológicas. La revolución cartesiana impregnó de epistemología todas las cuestiones filosóficas, creando con ello una radical distorsión al tiempo que una inestabilidad. En nuestros días se ha puesto de manifiesto esta confusión estableciéndose que una teoría de la conciencia es una teoría de una propiedad mental y como tal el discernimiento de su naturaleza es una cuestión metafísica en la que no deben inmiscuirse las preocupaciones que conciernen a la epistemología, como tampoco las que conciernen a su origen [15].

Podemos poner la cuestión metafísica de la siguiente manera: ¿Cuál es la constitución y/o estructura de la conciencia? ¿Cómo es que tenemos apariencias/experiencias? ¿Cómo surgen las apariencias de la estructura de la mente? ¿Cuál es la relación entre la propiedad de la conciencia y las

12. Consúltese J. Fodor (1987) y J. Searle (1986) sobre estos puntos.

13. Lo que tenemos por ahora es una perplejidad derivada de nuestra ignorancia, no sabemos si tiene solución o no la tiene y ningún argumento *a priori* nos constriñe en un sentido afirmativo o negativo.

14. Sobre este punto fascinante consúltese las dudas que aparecen en Churchland (1983 y 1984).

15. El más prominente defensor de esta distinción ha venido siendo J. Fodor. Véase Fodor (1991).

propiedades del pensamiento y de la intencionalidad? ¿Cuáles y cuántos son los mecanismos en que opera la conciencia? ¿Qué relaciones constitutivas o relacionales guarda la conciencia con otras propiedades? ¿Es la conciencia una propiedad simple o resulta de varios otros elementos constitutivos? ¿La conciencia fenomenal resulta de la conciencia representacional?

Una vez que se cuenta con respuestas para las cuestiones anteriores se está en mejor posición para responder acerca de los problemas de explicación, de conocimiento, etc. No se trata, sin embargo, de una cuestión de prioridades sino de una necesidad condicional en la que están involucradas las cuestiones semánticas, epistemológicas, científicas, morales, etc., con la cuestión metafísica.

La cuestión, en suma, reside en la posibilidad de lograr un nivel teórico de la propiedad de la conciencia, un nivel que exceda la experiencia ordinaria y tal vez tenga que introducir algo como una distinción apariencia-realidad, de lo que experimentamos frente a lo que es.

VII. TIPOS DE CONCIENCIA

Hemos introducido la denominación de conciencia fenomenal (CF) frente a la conciencia de... o conciencia representacional, proposicional o de contenido (CR). La CF se caracteriza por un percatarse subjetivo, por la sentiencia, la sensación, la apariencia, la experiencia. Sus propiedades son las de «la forma en que las cosas nos aparecen», «el carácter cualitativo», «el qualia», «las cualidades fenomenológicas inmediatas», «el como qué es ser x» (perspectival). Con esta denominación se hace referencia a las propiedades experienciales brutas, inmediatas, de las sensaciones, sentimientos y percepciones. Como tales, esas propiedades no son exclusivas de las personas sino también disfrutadas por los animales en formas y proporciones variadas.

Hay una enorme variedad de posiciones respecto de la naturaleza de la CF. Algunos piensan que tiene contenido, que ese contenido es representacional, pero no es intencional y tampoco es conceptual [16]. Si es verdad que en el centro de la CF hay una oscilación neural sincronizada en 35-75 hertz, no sabemos cómo tal oscilación puede constituir la base de la CF: no tenemos tan siquiera los conceptos que nos permitan explicar por qué tales oscilaciones están relacionadas con la CF. En suma, hay una perplejidad radical que nos impide comprender la naturaleza de la CF aun cuando tengamos experiencia cotidiana de ella y reconozcamos su importancia para nuestra sobrevivencia a la vez como individuos y como especie. Hay aquí lo que algunos autores califican como un hiato

16. Block (1993). Nagel (1974) introduce la cuestión más punzantemente general.

explicativo, a saber, que no hay investigación de la CF y que lo que parece constituir tal investigación es una tarea periférica tan «sólo una combinación de psicología cognitiva y exploración de síndromes neuropsicológicos» [17].

El otro tipo de conciencia, la conciencia representacional (CR), proposicional o de contenido es la que nos proporciona acceso a la información o contenido de los estados mentales. Este tipo de conciencia es representacional y juega un papel importante en el control racional de la acción y del lenguaje. Además, tiene una naturaleza funcional relativa a un sistema. En el modelo de Schakter de la CR, que aparece al final, un contenido se vuelve consciente debido a las relaciones informacionales entre el sistema ejecutivo y los otros módulos especializados.

Aun cuando existan dos tipos de conciencia CF y CR, un mismo contenido puede llegar a ser, ambos, CF (un destello de color rojo) y CR (proviene de un semáforo y representa una orden de detener el auto y esperar hasta que desaparezca y advenga el color verde). Un mismo contenido origina dos tipos de conciencias diferentes. La conciencia mínima que se da, por ejemplo, al despertarse (un caso de advertir), es un caso de CF, mientras que en el caso de la reflexión tenemos un ejemplo de CR. Sin embargo, hay casos en los que los dos tipos de conciencia ocurren separadamente y se pueden resolver varias cuestiones y perplejidades distinguiéndolos, como veremos más adelante.

VIII. LA METÁFORA COMPUTACIONAL

En nuestros días surge una gran promesa en las investigaciones de las propiedades mentales debido a la interpretación del funcionamiento de la mente en términos de un ordenador o computadora. La idea es que la mente es una máquina sintáctica que opera como una máquina de tipo Turing, con entradas y salidas de información y módulos que llevan a cabo diversos procesos. Esos módulos se introducen como cajas negras de las que se ignora su constitución, pero se las va conociendo conforme se descubre el mecanismo operativo. El modelo de Schackter (reproducido al final) ejemplifica esta teoría funcional y computacional; dicho modelo tiene que interpretarse de acuerdo con la distinción anterior de CF y CR, pues en una interpretación la CF resulta una condición necesaria de todo otro tipo de conciencia, mientras que en otra interpretación la CR ocupa la posición de condición necesaria y la CF queda relegada a una condición de mero reflejo. Pero ¿cómo puede ayudar un modelo cogni-

17. Block (1993). Podemos distinguir entre un problema científico que es el de dar una explicación de la propiedad de la conciencia y el problema filosófico que consiste en dar cuenta de como qué sería ser aparecido por *x*.

tivo a resolver las dificultades específicas de la propiedad de la conciencia? Consideremos un par de casos a manera de ilustración con tres modelos diferentes.

IX. TRES MODELOS COGNITIVOS: DENNETT, ROSENTHAL, SCHAKTER

Hagamos una breve descripción de los modelos cognitivos de Dennett, Rosenthal y Schakter que aparecen al final. En el modelo de Dennett la información que ingresa por los sentidos se ve procesada por un complejo de cajas negras de control, memoria y solución de problemas para desembocar en lenguaje y las subrutinas motoras de la acción en el mundo. Este modelo se ve alterado por las tesis de Rosenthal (1986) [18] y del propio Dennett (1991).

Rosenthal objeta que el modelo de Dennett no describe apropiadamente la función específica de la conciencia y él propone que se trata de una capacidad que tienen algunos estados mentales (EM) de disparar un pensamiento inconsciente P1 que incide sobre el EM que lo disparó y debido a esta incidencia lo vuelve consciente. P1 es un pensamiento de una jerarquía más alta que EM y por ello se le puede llamar a esta teoría del pensamiento de orden superior (POS). Esta incidencia puede repetirse sobre el P1 por P2 que entonces vuelve consciente a P1 y así sucesivamente según la capacidad de cada persona. Esta incidencia es un hecho contingente cuya especificación y ejemplificación tienen que descubrirse empíricamente. La incidencia de actos de pensamiento sobre EM permite, de acuerdo con Rosenthal, explicar tanto los casos normales como los casos difíciles redescribiendo el darse cuenta, el percatarse, la atención, la reflexión, la reflexión de segundo grado, etc., que resultan así estadios particulares de una misma propiedad general de la conciencia.

Dennett mismo se autocritica también por la falta de especificidad de su primera teoría y suplementa su modelo con una teoría posterior de los múltiples esbozos de acuerdo con la cual lo específico de la propiedad de la conciencia es elaborar varios esbozos de lo que sucede en la mente, sin que ninguno de ellos sea definitivo o único sino aproximativo, intentando ofrecer la versión más fiel de lo acontecido. Todo esbozo es revisable y lo que decide si un estado mental es consciente es que llegue a la memoria y se manifieste en la conducta del individuo. Fuera de estos criterios no hay un hecho al cual se pueda apelar para decidir si algún estímulo arribó a la conciencia o no y ciertamente la introspección carece de todo valor en este respecto. Esta tesis de Dennett puede verse como una versión de la teoría de la interpretación aplicada a la teoría del POS.

18. Rosenthal se inspira en Brentano y en Armstrong (1968) al elaborar su teoría.

Finalmente en el modelo de Schakter la información ingresa a través de módulos especializados los cuales la envían a un sistema ejecutivo mediado por un sistema de memoria, un sistema de respuesta y hasta arribar a un sistema de procedimiento. El modelo puede tener dos lecturas [19], a saber, una en que la CF es epifenomenal y otra en la que no lo es. De acuerdo con la primera, una vez que la información recibe la elaboración en los módulos y sistemas el sistema ejecutivo la manda al sistema de procedimiento y a la CF simultáneamente de manera que la persona se da cuenta del resultado sin que este darse cuenta tenga nada que ver con la constitución y elaboración del mismo; de acuerdo con la segunda, la CF interviene en la elaboración de la información y conjuntamente con los módulos y sistemas arriba a un resultado que envía al sistema ejecutivo y éste actúa sobre el sistema de procedimiento. Aquí de nuevo se trata de relaciones contingentes cuya implementación e instanciación deben descubrirse mediante la investigación empírica.

X. LOS CASOS DIFÍCILES

Los casos difíciles resultan de una crucial importancia para fijar la naturaleza de la conciencia, pues ellos exhiben perplejidades y constriñen el tipo de experimentos que necesitamos idear para encontrar la solución de las dificultades y con base en esa solución se va descubriendo la naturaleza de la propiedad [20]. Tomemos un par de casos, uno es el del escucha dicótico y el otro es el de percepción visual de dos datos en sucesión temporal.

En el primer caso, una persona escucha por el oído izquierdo la información I mientras que simultáneamente por el oído derecho se le suministra la información D (que tiene que ver con el contenido de la información I) pero de tal manera que la persona no la advierte (si se le pregunta, ella niega haber recibido la información D). Sin embargo, al resolver cuestiones que tienen que ver con ambas informaciones I y D la persona utiliza la información D (que niega haber recibido y guardar en su memoria) para resolver esas cuestiones. Este primer caso refuta la igualdad Cartesiana (mente = conciencia), pero, más importante aún, revela que la actividad del pensamiento opera detrás del umbral de la conciencia y que por lo tanto la conciencia no constituye una condición necesaria de esos procesos del pensamiento. La imagen que dibuja es la de una corriente de EM en la que hay picos que rebasan el umbral de la conciencia y esos picos son los EM conscientes; la mayor parte de los procesos y EM transcurren en la no-conciencia. El caso sugiere una pri-

19. Block (1993) propone estas dos lecturas del modelo de Schakter.

20. No puedo describir aquí las relaciones entre estos niveles o estadios teóricos que son de la mayor importancia.

mera cuestión crucial, a saber, hasta qué punto la conciencia es una propiedad primaria de la mente o se trata solamente de una propiedad secundaria, derivada de la interacción de otras propiedades mentales.

El segundo caso se trata de un metacontraste en que a una persona se le suministra una breve presentación (30 milisegundos) de un disco seguido inmediatamente por una presentación de una dona cuyo borde interno ocurre justo donde apareció el borde externo del disco. La persona niega haber percibido el disco y solo reconoce la dona, pero hay evidencia de que la persona tiene información acerca del disco.

XI. LA ADECUACIÓN DE LOS MODELOS COGNITIVOS

Evaluemos brevemente la forma en que los tres modelos cognitivos resuelven los casos difíciles. De acuerdo con el primer modelo de Dennett el caso del escucha dicótico se resuelve porque la información I y D ingresa y recibe todo el procesamiento, pero la información D no accede del todo al módulo del control-atención y sin embargo se encuentra allí y se manifiesta en la verbalización. Rosenthal encuentra ese modelo de Dennett insuficientemente discriminatorio y necesitado de reformas que permitan esclarecer cómo sucede que en un caso la información se torna consciente y en el otro no. Con la enmienda de Rosenthal esta perplejidad desaparece, pues en un caso la información I dispara un pensamiento de orden superior que hace que esa información aparezca o se advierta, mientras que en el otro caso la información D no dispara ese pensamiento y aun cuando ingresó y es procesada en los módulos —para reaparecer después— no queda advertida en el momento de su ingreso por el escucha dicótico.

En el modelo posterior de Dennett las cosas se complican, pues la información sin duda ingresa a los módulos perceptuales pero permanece oculta para la verbalización (consciente). No parecen aplicarse los múltiples esbozos, pues la naturaleza del caso establece que ambas informaciones arribaron, pero una no quedó registrada y sin embargo de no dejar registro, se manifiesta en la conducta posterior del agente integrada con la información I cuyo ingreso sí quedó registrado. Aquí presumiblemente un primer esbozo debe negar el ingreso de la información D y posteriormente un esbozo diferente debe aceptar que esa información sí arribó (inconscientemente) y está guardada en la memoria del sujeto.

El modelo de Shakter adaptado por Ned Block implica un cambio sustancial, pues por una parte exhibe la entrada de información a los módulos especializados y por la otra expresa el paso de los módulos a la conciencia fenomenal, la cual desempeña un papel crucial en el sistema ejecutivo mandando las directivas a la acción. Aquí la CR y la CF interactúan y coadyuvan en la acción resultante. Hay una diferencia entre los

dos tipos de conciencia, pero ambas acaecen y resultan en la acción. Sin embargo, hay dos posibilidades en el caso del escucha dicótico, una es afirmar que la información D pasó por los módulos pero no llegó a la CF; la otra es afirmar que sí alcanzó la CF —pues es indispensable— pero no se manifestó verbalmente.

El caso del metacontraste con dos informaciones percibidas visualmente presenta un problema para el primer modelo de Dennett, pues no discrimina suficientemente para poder ofrecer una solución de por qué el escucha solamente reporta una información y suprime la otra. En su modelo posterior Dennett ofrece una solución en términos de dos bosquejos, uno Estaliniano y otro Orwellesco: según el primero, la información del disco es editada antes de arribar a la conciencia y nunca llega a ella; de acuerdo con el segundo, la información del disco sí llega a la conciencia pero queda suprimida subsecuentemente del recuerdo. La única información que llega a la conciencia y es retenida en la memoria es la información de la dona y esto lo sabemos porque se la recuerda y porque se manifiesta en la conducta verbal y física del agente, no por ningún hecho privado o externo.

Rosenthal piensa que su teoría POS puede resolver este caso diciendo que ambas, la información de la dona y la del disco llegan a la mente, pero solamente la primera se torna consciente debido a la incidencia de un pensamiento de orden superior; la información del disco permanece en el recuerdo sin manifestarse verbalmente pero apareciendo parcialmente en la conducta del agente (y en esa medida hay que suponer que hubo algún pensamiento, con alguna fuerza, que incide sobre esa información alojada en la memoria).

En el modelo de Schakter interpretado por Block hay CF de ambos, el disco y la dona, pero solamente esta última alcanza a la verbalización y a la CR; sin embargo, ambas informaciones se manifiestan en la conducta del agente aun cuando de maneras diferentes y por lo tanto con diversos grados de CR.

XII. CIENCIA COGNITIVA DE LA CONCIENCIA: ESCEPTICISMO Y OPTIMISMO

Hay dos tipos de escepticismo: uno niega que la conciencia exista y sea una propiedad de las personas y por lo tanto del mundo[21]. El otro escepticismo afirma que la propiedad de la conciencia existe pero niega, conjuntamente con el escepticismo anterior, que pueda haber un conocimiento científico de esa propiedad de la conciencia[22]. Correlativamen-

21. Esta es la posición de Churchland (1983) y los eliminativistas. Los epifenomenalistas según el tipo de que se trate comparten este escepticismo en diferentes grados. Piénsese sobre todo en el caso de la CF.

22. Esta es la posición de Nagel (1974) y Searle (1987 y 1991). McGinn (1991) lleva este

te, las posiciones optimistas afirman indistintamente que la conciencia existe y/o que puede haber un conocimiento del mismo tipo que el de la ciencia física [23].

XIII. REALISMO ACERCA DE LA PROPIEDAD DE LA CONCIENCIA

Esta última consideración desemboca en la cuestión de la realidad ontológica que tiene la propiedad de la conciencia: ¿se trata de una propiedad con existencia propia, autónoma, o de una propiedad que solamente existe en otras propiedades como un resultado de ellas como algo emergente o como algo que sobreviene de una base firme? ¿Qué tanta realidad ontológica tiene la propiedad de la conciencia (en particular la CF)? ¿En un inventario de lo que existe se tiene que incluir a la conciencia? ¿Está la propiedad de la conciencia entre las propiedades básicas del mundo o solamente entre las accidentales y dispensables? Las teorías que alimentan los modelos cognitivos anteriores, por ejemplo, sostienen diversos grados de realismo de la propiedad de la conciencia. Los modelos de Dennett la vuelven una cuestión de interpretación con elementos contextuales, genéticos y sociales. El modelo de Rosenthal la vuelve contingente sobre la incidencia de un pensamiento de orden superior y la tesis de Block afirma su realidad irrestricta.

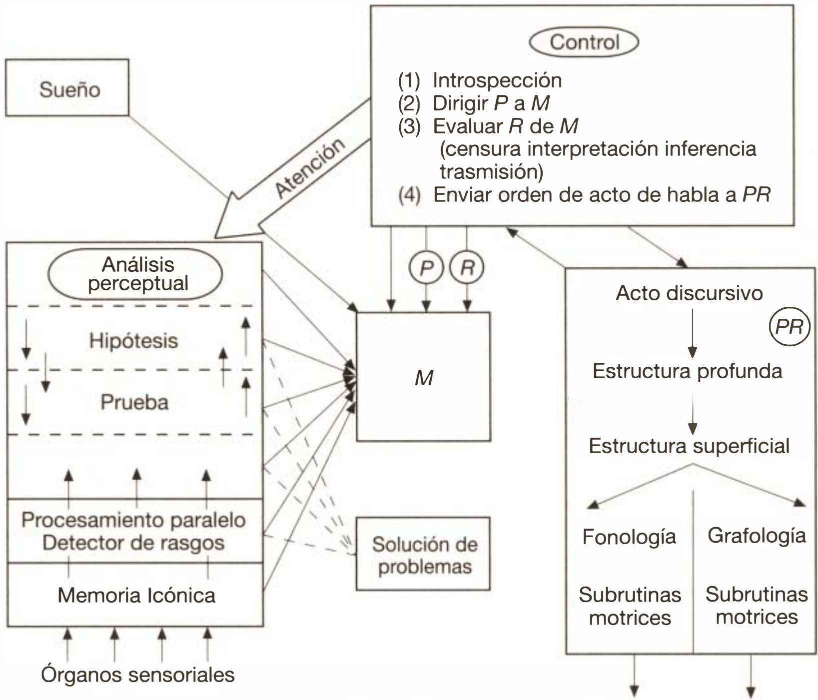
XIV. LAS PERSONAS Y LA CONCIENCIA

Terminemos con una última cuestión: ¿qué lugar ocupa la conciencia en el concierto de propiedades que constituyen la copersonalidad, es decir, la unidad sincrónica de la persona y por ello mismo su identidad? [24]. En todas las teorías que hemos venido examinando la conciencia resulta una propiedad indispensable de las personas, pero mientras Rosenthal piensa que se la puede investigar por separado, como una parte de la teoría del pensamiento, Dennett y Block parecen inclinarse por su estudio en conjunción con otras propiedades mentales, no aisladamente, sino en conjunto. Para estos últimos tener una teoría de la conciencia es tener una teoría de la mente; el estudio de la conciencia no puede desmembrarse del estudio de muchas otras propiedades mentales.

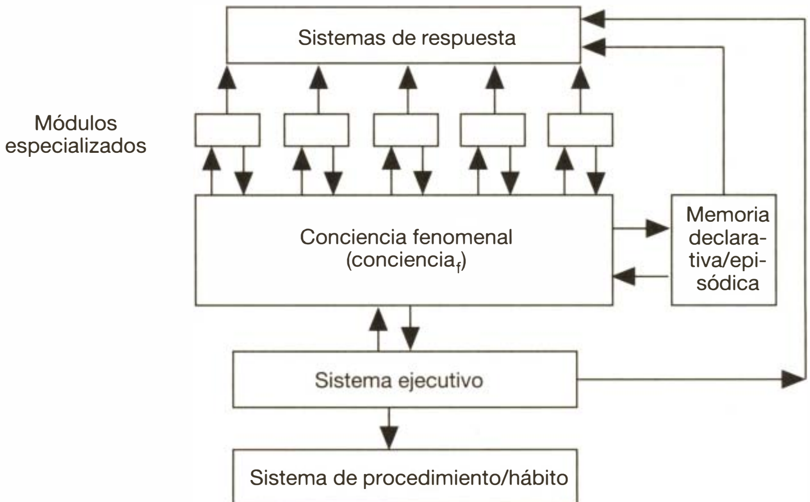
punto de vista a un extremo y sostiene que aun cuando es un hecho contingente, nunca podremos conocer la propiedad de la conciencia y por ello dicha propiedad tiene el carácter de un *noumeno*.

23. Estas posturas optimistas, en la medida en que rechazan que la conciencia sea una propiedad no-física, se las califica hoy día de naturalistas. Sin embargo hay muchas tesis diferentes encubiertas bajo este adjetivo.

24. Sobre estas nociones de copersonalidad y de unidad sincrónica e identidad consúltese mi libro *Las Personas*, en prensa.



Modelo de Daniel Dennett (tomado de *Brainstorms*, p. 155)



Modelo de Ned Block (tomado de BBS, 1991, 14:4). En este modelo la conciencia$_f$ juega un papel causal, no es epifenoménica.

XV. CONCLUSIONES

Hemos llevado a cabo un recuento histórico selectivo de los problemas teóricos que suscita la propiedad de la conciencia. Hemos visto que la tendencia predominante se aparta del Cartesianismo. Hemos reparado en la perspectiva contemporánea; hemos visto, en particular, la esperanza que representan los modelos cognitivos. Sin embargo, advertimos que no hay aún una concepción de lo que puede constituir una solución para el problema filosófico que ocasiona la CF; el problema empírico de la explicación, tal vez la tenga —y aquí la principal esperanza puede estar en la investigación cognitiva, pero aun si alcanzara ese tipo de solución, resulta difícil pensar que las personas, tal como están constituidas, dejaran de experimentar la perplejidad que experimentan con la CF y que ésta no constituyera una propiedad metafísica problemática.

BIBLIOGRAFÍA

Armstrong, D. M. (1968), «What is consciousness?», en *The Nature of Mind*, Cornell University Press, Ithaca.
Block, N. (1980), «Troubles with Functionalism», en Block (ed.), *Readings in the Philosophy of Psychology*, MIT Press, Cambridge, Mass.
Block, N. (1990), «Consciousness and accessibility»: *Behavioural and Brain Sciences, 13*, 596-598.
Block, N. (1992), «Begging the question against phenomenal consciousness»: *Behavioural and Brain Sciences*.
Block, N. (1993),«On a confussion about a function of consciousness»: (forthcoming): *Behavioural and Brain Sciences*, y en J. L. Díaz y E. Villanueva (eds.), *La Conciencia*.
Brentano, F, (1874), *Psychologie vom empirischen Standpunckt*, Leipzig
Churchland, P .S. (1983), «Consciousness: the transmutation of a concept»: *Pacific Philosophical Quarterly*, *64*, 80-93
Churchland, P. S. (1984), *Matter and Consciousness*, MIT. V.e.: *Materia y Conciencia,* Barcelona, Gedisa.
Crick, F. y Koch, C. (1990), «Towards a neurobiological theory of consciousness»: *Seminars in the Neurosciences,* 2, 263-275.
Davies, M. y Humphreys, G. (eds.) (1993), *Consciousness*, Blackwell, London.
Descartes, R. (1897), *Oeuvres*, Adam et Tannery, Paris
Dennett, D. (1978), *Brainstorms*, MIT, Cambridge Mass.
Dennett, D. (1991), *Consciousness Explained*, Little Brown, New York.
Díaz, J. L. y Villanueva, E. (eds.) (1994), *La Conciencia*, FCE, México.
Dretske, F. (1993), «Conscious Experience»: *Mind, 102*, 406, 263-284.
Fodor, J. (1983), *The Modularity of the Mind*, MIT Press, Cambridge, Mass.
Fodor, J. (1991), *A Theory of Content*, MIT Press, Cambridge, Mass.
Goldman, A. (1993), «Consciousness, folk psychology and cognitive science» (forthcoming): *Consciousness and cognition*.
Loar, B. (1990), «Phenomenal Properties», en J. Tomberlin (ed.), *Philosophical*

Perspectives: Action Theory and Philosophy of Mind, Ridgeview, Atascadero, Cal.
Jackson, F. (1986), «What Mary didn´t know»: *Journal of Philosophy, 83,* 291-5
James, W. (1890), *The Principles of Psychology*, Holt.
Mc Ginn, C. (1991), *The Problem of Consciousness*, Blackwell, London.
Marcel, A. J. y Bisiach, E. (eds.) (1988), *Consciousness in Contemporary Science*, Oxford University Press, Oxford.
Nagel, T. (1974), «What is it like to be a bat?»: *Philosophical Review.*
Nagel, T. (1979), *Mortal Questions*, Cambridge University Press, Cambridge.
Nagel, T. (1986), *The View from Nowhere,* Oxford University Press, Oxford.
Rosenthal, D. (1986), «Two Concepts of Consciousness»: *Philosophical Studies, 49*, 329-359.
Rosenthal, D. (1993), «The higher-order thought theory of consciousness», en J. L. Díaz y E. Villanueva (eds.), *La Conciencia.*
Schakter, D. (1989), «On the relation between memory and consciousness: dissociable interactions and conscious experience», en H. Roediger y F. Craick (eds.), *Varieties of Memory and Consciousness: Essays in Honour of Endel Tulving*, Erlbaum, Hillsdale, NJ.
Searle, J. (1987), «Consciousness, explanatory inversion and cognitive science»: *Behavioural and Brain Sciences, 13*, 4, *585-595*
Searle, J. (1991), «Consciousness, Unconsciousness and Intentionality», en E. Villanueva (ed.), *Consciousness*, 1991.
Villanueva, E. (1989), *Ensayos de Historia Filosófica*, UNAM, México.
Villanueva, E. (ed.) (1992), *Consciousness*, Ridgeview, Atascadero, Cal.
Villanueva, E. (1994), «Ciencia Cognitiva y Conciencia»: *Contextos.*
Villanueva, E. (1994), *Las personas*, en prensa.

ÍNDICE ANALÍTICO *

* El criterio seguido para la selección de entradas en este índice ha sido el de señalar únicamente aquellas páginas donde se desarrollan los conceptos indicados por la entrada léxica.

ÍNDICE DE NOMBRES *

* El criterio seguido para la selección de entradas en este índice ha sido el de señalar únicamente las páginas en las que se expresan ideas o argumentos relevantes del autor citado.

NOTA BIOGRÁFICA DE AUTORES

Juan José Acero (Madrid, 1948), catedrático de lógica de la Universidad de Granada, es especialista en filosofía del lenguaje. Coautor de *Introducción a la filosofía del lenguaje* (1982), destacan entre sus obras *Filosofía y análisis del lenguaje* (1985) y *Filosofía y lenguaje* (1993).

Fernando Broncano (Salamanca, 1954), es profesor de lógica y filosofía de la ciencia en la Universidad de Salamanca. Su labor investigadora se centra en la filosofía de la ciencia y de la técnica. Entre sus publicaciones destacan *Metaciencia, falibilismo y racionalidad* (1981) y su aportación a la obra colectiva *Perspectivas actuales en Lógica y Filosofía de la Ciencia* (1994). Es editor de *Nuevas meditaciones sobre la técnica*, de próxima aparición en esta Editorial.

Josep Corbí (Monóvar, Alicante, 1957), ejerce la docencia en el Departamento de Metafísica y Teoría del Conocimiento de la Universidad de Valencia, donde es un reconocido especialista en la filosofía de la mente y teoría de la racionalidad. Es autor de numerosos artículos, publicados en revistas especializadas y obras colectivas.

Jesús Ezquerro (Autol, La Rioja, 1951), profesor de lógica y filosofía de la ciencia en la Universidad del País Vasco, centra su interés en la filosofía del lenguaje y ciencia cognitiva. Es secretario del Comité de Programación del International Colloquium on Cognitive Science (ICCS). Entre sus publicaciones cabe destacar *Cognition, Semantics and Philosophy* (1992) y *Philosophy and Cognitive Science: Categories, Consciousness and Reasoning* (1995).

Alfonso García Suárez (La Felguera, Oviedo, 1948), especialista de filosofía del lenguaje en la Universidad de Oviedo, ha publicado, entre otros trabajos, *La lógica de la experiencia: Wittgenstein y el problema del lenguaje privado* (1976) e *Interpretaciones del pensamiento en Wittgenstein* (1991).

Manuel García-Carpintero (Daimiel, Ciudad Real, 1957), es profesor de filosofía del lenguaje en la Universidad de Barcelona. Entre sus publicaciones destacan sus contribuciones a obras colectivas y sus numerosos artículos en revistas especializadas.

Antoni Gomila Benejam (Ciutadella, Baleares, 1963), experto en filosofía de la mente y filosofía de la psicología, desarrolla su labor docente en el Departamento de Filosofía de la

Universidad de las Islas Baleares. Es autor de numerosos artículos en revistas y obras especializadas.

Manuel Liz (Burgos, 1960), es profesor en la Facultad de Filosofía de la Universidad de La Laguna y especialista en lógica y filosofía de la ciencia. Amén de sus numerosas aportaciones a revistas especializadas, destaca su obra *La vida mental de algunos trozos de materia. Teorías de la sobreveniencia de lo mental.*

Josep L. Prades (1954), profesor del Departamento de Filosofía de la Universidad de Murcia, su actividad investigadora se centra en la filosofía de la mente y la teoría del conocimiento. Ha colaborado en la obras colectivas *En torno a Wittgenstein* (1993) y *Mirar con cuidado* (1994), y es autor de *Wittgenstein: Mundo y Lenguaje* (1990).

Daniel Quesada (Barcelona, 1947), es especialista en lógica y filosofía de la ciencia, disciplina que imparte en la Universidad Autónoma de Barcelona. Autor de *La Lógica y su Filosofía* (1985), así como de numerosos artículos y colaboraciones en obras colectivas y revistas especializadas.

Eduardo Rabossi (Buenos Aires, 1930), especialista en ética, filosofía de la mente y filosofía de los derechos humanos, es catedrático de filosofía del derecho en la Universidad de Buenos Aires. Es autor, entre otras, de *Análisis filosófico, lenguaje y metafísica* (1977) y *Estudios éticos, cuestiones conceptuales y metodológicas* (1979), así como editor de la obra *Ética y análisis* (1985).

Vicente Sanfélix Vidarte (Valencia, 1957), desarrolla su actividad académica en la Universidad de Valencia y es un experto en epistemología, filosofía moderna y filosofía actual. Es autor de *Hume. Textos Cardinales* (1986), coautor de *Wittgenstein: Mundo y Lenguaje* (1990) y editor de *Acerca de Wittgenstein* (1993).

Josefa Toribio Mateas (Santiago de Calatrava, Jaén, 1961), es profesora asistente de filosofía en el Departamento de Filosofía de la Washington University in St. Louis. Es autora de numerosos artículos publicados en revistas especializadas.

Enrique Villanueva (México), presidente de la Sociedad Filosófica Iberoamericana de Filosofía y especialista en filosofía de la mente, es investigador del Instituto de Investigaciones Filosóficas, dependiente de la Universidad Nacional Autónoma de México. Editor de *Information semantics and Epistemology* (1990), es autor además de *El argumento del Lenguaje Privado* (1985) y *Ensayos de Historia Filosófica* (1989).